Universidad del Tolima

FACULTAD DE CIENCIAS

DEPARTAMENTO DE MATEMÁTICAS Y ESTADÍSTICA

ESTUDIO DEL IPC EN EL DEPARTAMENTO DEL TOLIMA Y DEPARTAMENTOS ALEDAÑOS MEDIANTE GAMs

Trabajo de grado para optar al título de Especialista en Estadística

Jorge Edgar Silva Veloza

Director:

Mg. Alfonso Sánchez Hernández

Departamento de Matemáticas y Estadística

Especialización en Estadística

POSTGRADOS

IBAGUÉ

2018

ESTUDIO DEL IPC EN EL DEPARTAMENTO DEL TOLIMA Y DEPARTAMENTOS ALEDAÑOS MEDIANTE GAMs

Autor:
Jorge Edgar Silva Veloza
Código:
0000001122

Trabajo de grado para optar al título de Especialista en Estadística

Director:
Alfonso Sánchez Hernández
MSc. en Investigación Operativa y Estadística

UNIVERSIDAD DEL TOLIMA
FACULTAD DE CIENCIAS
DEPARTAMENTO DE MATEMÁTICAS Y ESTADÍSTICA
IBAGUÉ
2018

UNIVERSIDAD DEL TOLIMA

FACULTAD DE CIENCIAS

ESPECIALIZACIÓN EN ÉNFASIS EN ESTADÍSTICA

ACTA DE SUSTENTACIÓN TRABAJO DE GRADO

TÍTULO: ESTUDIO DEL IPC EN EL DEPARTAMENTO DEL TOLIMA Y DEPARTAMENTOS ALEDAÑOS MEDIANTE GAMs

AUTORES: Jorge Edgar Silva Veloza (código 0000001222)

DIRECTOR: ALFONSO SANCHEZ HERNÁNDEZ

JURADOS: YURI MARCELA GARCIA SAAVEDRA
JAIRO ALFONSO CLAVIJO

CALIFICACIÓN: 4.6 (CUATRO SEIS) MERITORIO

X APROBÓ ___ REPROBÓ

OBSERVACIONES:

FIRMAS:

YURI MARCELA GARCÍA S.
Jurado 1

JAIRO ALFONSO CLAVIJO
Jurado 2

ALFONSO SANCHEZ HERNÁNDEZ
Director del Trabajo

LEONARDO DUVAN RESTREPO
Secretario Académico

Ciudad y fecha: Ibagué, 5 de julio de 2018

Índice General

Índice de tablas

Índice de figuras

RESUMEN

Los Modelos Aditivos Generalizados comúnmente conocidos como GAMs fueron propuestos por Hastie y Tibshirani (1986, 1990 [10]). Un GAM es simplemente un Modelo Lineal Generalizado, en el que intervienen uno o varios predictores, los cuales involucran una suma de funciones suaves de covariables. La variable respuesta Y_i pertenece a alguna distribución de la familia exponencial, mientras que la parte estrictamente paramétrica en el predictor del modelo se mantiene. Las covariables que se expresan en términos de funciones suaves, son variables que pueden ser indexadas en tiempo ó espacio, éstas a su vez pueden ser expresadas mediante combinaciones lineales de funciones suaves (tipo Fourier ó B-Splines). Algunos autores llaman a este tipo de modelos *semiparamétricos*. Además se pueden considerar modelos en donde sólo intervienen covariables de este tipo, las cuales se llaman semiparamétricas. El indice de precios al consumidor en un país o en un departamento, no puede escapar a la aplicabilidad de este tipo de modelos, por tal razón el presente trabajo intenta realizar una aplicación estadística mediante modelamiento al Indice de Precios al Consumidor (IPC) en el Tolima y departamentos circunvecinos.

INTRODUCCIÓN

El Indice de precios al consumidor (IPC) es la variación absoluta o porcentual en los precios de la canasta familiar, la cual está definida por el Departamento Nacional de Estadística (DANE) como un índice de canasta final, correspondiente a un periodo base de tiempo, construido sobre una variante de los índices tipo *Laspeyres*, que permite una actualización más rápida de la canasta para seguimiento de precios, según evolucione o cambie el gasto de consumo de los hogares de un país. Esta cantidad puede variar a nivel regional y por lo tanto en su acumulado nacional. A partir de 2010 se viene realizando en las capitales de los 23 departamentos y también se consolida un IPC a nivel nacional. Esta variación puede ser distinta en las diferentes regiones del país.

Teniendo en cuenta que la base fundamental de la estadística es el estudio de la variabilidad en espacio o tiempo, el presente trabajo pretende realizar una comparación del IPC en el Tolima y los departamentos circunvecinos (Cauca, Valle, Quindío, Risaralda, Caldas, Cundinamarca y Huila). Se tomará el IPC de cada una de las ciudades capitales en las que el DANE mide la variación.

Sabiendo que el IPC es una variación porcentual se pretende implementar una metodología estadística basada en Modelos Lineales Generalizados (GLM), más especificamente en Modelos Aditivos Generalizados (GAMs), la cual relaciona una variable respuesta Y, en (0,1), en nuestro caso el IPC para el departamento del Tolima con otras variables independientes : X_1 el IPC para los demás departamentos aledaños al Tolima, y de ser posible otras variables independientes X_2, la distancia en Kilómetros entre las ciudades capitales, X_3 la tasa de desempleo de cada una de las ciudades capitales, etc.

Una vez determinado el modelo se realizará un análisis de sensibilidad del mismo, y se realizará un diagnóstico con el fin de determinar posibles valores influenciales, atípicos o outliers.

Posteriormente los resultados del trabajo de investigación se socializarán en un evento academico de caracter regional o nacional.

PLANTEAMIENTO DEL PROBLEMA

Con base en lo anterior, se sugieren las siguientes preguntas de investigación:

- ¿Se puede establecer algún modelo funcional que relacione el IPC del Tolima con el IPC de los departamento aledaños?
- ¿Existirá alguna variacion ó correlación entre el IPC del Tolima y los Departamentos aledaños?
- ¿Existe alguna diferencia sustancial en el comportamiento del IPC en el departamento del Tolima, frente a los departamentos vecinos?
- ¿Se pueden identificar algunas variables independientes, que incidan significativamente en el IPC en los departamentos del centro del país?
- ¿Se podrán identificar factores de naturaleza fija o aleatoria que afecten funcionalmente al IPC como variable dependiente?
- ¿Podrá utilizarse la estadistica funcional en el modelo propuesto para incluir variables de tipo continuo que permitan dilucidar comportamientos extraños, referentes al IPC?

JUSTIFICACIÓN

En cualquier región del país surgen problemáticas de índole social, que afectan de manera directa o indirecta a la población de la región y por ende al país. La calidad de vida manifiesta de cada región está completamente asociada a la productividad, a los recursos naturales a la explotación de los mismos. Anteriormente algunas regiones del país, debido a su desarrollo, se encontraban estipuladas como territorios nacionales, tal era el caso de las Intendencias y Comisarías. Esta situación permitía apreciar una gran diferencia social entre los departamentos y los territorios nacionales, por ende el presupuesto de la nación era limitado para esas regiones. A partir del proceso de descentralización, es decir la conformación de los departamentos, estas regiones no trajeron consigo el desarrollo esperado , sino que por el contrario se vieron afectadas por la presencia de grupos al margen de la ley. Como se evidencia el pleno desarrollo industrial y agroindustrial se siguió centralizando en las más importantes ciudades capitales del país. Se puede decir que el Indice de Precios al Consumidor es uno de los indicadores de la economía más relevantes y más precisos para medir la calidad de vida en cada una de las regiones. No obstante, los departamentos ubicados en el centro, norte y occidente del país han sido los más beneficiados por su conexión terrestre con los puertos marítimos. Se aprecia también que el departamento del Tolima y las regiones circunvecinas son privilegiadas por su cercania a la capital y por ser corredores terrestres en menor medida fluviales, para la transferencia de los bienes de consumo. En este orden de ideas, para el presente trabajo se implementa una metodología estadística basada en modelos lineales generalizados (GLM), para explicar el comportamiento del Indice de Precios al Consumidor (IPC) en el departamento del Tolima, con respecto a las regiones circunvecinas.

OBJETIVOS

OBJETIVO GENERAL

Estudiar el comportamiento de la variación del Índice de Precios al Consumidor (IPC) en el departamento del Tolima, con respecto a los departamentos circunvecinos entre los años 2010-2017

OBJETIVOS ESPECIFICOS

- Implementar una metodología estadistica, basada en Modelos Lineales Generalizados (GLM), para el estudio de la variación del Índice de Precios al Consumidor (IPC) usando la información del departamento del Tolima y departamentos circunvecinos.
- Determinar algunos modelos lineales que permitan medir la variación del IPC Tolima frente al IPC de departamentos circunvecinos.
- Estimar los parámetros poblacionales asociados a los modelos encontrados.
- Realizar un estudio de sensibilidad e influencia local para realizar la validación del modelo estimado.
- Interpretar de una manera estadística el comportamiento de la información relacionada al IPC en el Tolima y departamentos circunvecinos, con el fin de pronosticar comportamientos futuros del IPC, en esta región del país.

METODOLOGÍA

Con el propósito de implementar una metodología estadística basada en modelos lineales generalizados (GLM), más especificamente, Modelos Aditivos Generalizados (GAMs) para el estudio de la variación del Índice de precios al consumidor (IPC), usando la información del IPC del Tolima y departamentos aledaños, se consultó la página del Departamento Nacional de Estadística (DANE), la información del IPC para el Departamento del Tolima y los siete departamentos a su alrededor, que están siendo publicados desde el año 2010. Con esta información se determinarán las variaciones en el tiempo del IPC para nuestro objeto de estudio. Posteriormente se representarán los parámetros poblacionales de la variación del IPC y se calculará un modelo estadístico para relacionar el comportamiento del IPC en el Tolima y los departamentos circunvecinos. Para terminar, se realizará un estudio de sensibilidad basado en diagnóstico para el modelo estimado y se interpretará el comportamiento de la información, con el fin de pronosticar el IPC del Tolima en un tiempo relativamente corto.
Por último se pretende evidenciar mediante algunos estadísticos sencillos de carácter funcional, examinar la posibilidad de que en estudios posteriores se aplique la recién descubierta herramienta estadística denominada *Datos Funcionales*, que en los últimos años ha sido aplicada en diferentes áreas del conocimiento, especificamente en el campo económico.

1. REFERENTES TEÓRICOS

1.1. Modelos Lineales Generalizados (GLM)

Los Modelos Lineales Generalizados (GLM) fueron propuestos por Nelder y Wedderburn (1972, [17]), se crearon a partir del gran desarrollo computacional después de los años 70 y algunos modelos que exigían procesos iterativos para la estimación de sus parámetros, comenzaron a tener mayor utilidad. Por ejemplo el modelo normal no lineal, el cual asume una estructura no lineal en los parámetros tuvo un gran avance, Paula (2013, [19], pp. 1). Gran cantidad de textos especializados en este tópico han sido escritos, dentro de ellos se pueden citar: Nelder y McCullag (1989, [16]), Dobson (2000, [7]) y Agresti (2017, [1]) entre muchos otros.

1.1.1. Componentes de un Modelo Lineal Generalizado

Un modelo lineal generalizado está formado por tres partes esenciales:

- **Componente aleatoria:** representa la variable respuesta Y, que puede ser de valor real ó un vector $\mathbf{Y} = (\mathbf{Y_1}, \mathbf{Y_2}, \cdots . \mathbf{Y_n})$. Cada variable ó componente del vector pertenece a la familia exponencial.
- **Componente sistemática:** corresponde al predictor lineal del modelo y se nota $\eta_i = \mathbf{X_i}^T \boldsymbol{\beta} = \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots , \beta_p X_{ip} = \sum_{j=1}^{p} X_{ij}\beta_j$, para la componente i-ésima del vector $\mathbf{Y}$, ó $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$ para el vector completo $\mathbf{Y}$.
- **Función de enlace o link:** relaciona la esperanza matemática de la variable dependiente con el predictor lineal, $g(\mu_i) = \eta_i$ para $i = 1, 2, \cdots , n$. La función $g(\cdot)$ debe ser monótona, suave y diferenciable.

Una distribución pertenece a la familia exponencial, si su función de densidad de probabilidad (variable aleatoria contínua) ó función de densidad discreta (variable aleatoria discreta) se puede escribir como:

$$f(y_i, \theta_i, \phi) = exp\big[\phi\{y_i\theta_i - b(\theta_i)\} + c(y_i, \phi)\big]$$

donde b y c son funciones arbitrarias, ϕ es un parámetro de dispersión y θ_i es conocido como el parámetro canónico de la distribución.

$$E(Y_i) = \mu_i = b'(\theta_i) \quad \text{y} \quad var(Y_i) = \phi^{-1} V_i; \quad V_i = \frac{\partial \mu_i}{\partial \theta_i}$$

A V_i se le conoce como función varianza, también se nota $V(\mu_i)$.
Un caso particular ocurre cuando el parámetro canónico θ_i coincide con el predictor lineal η_i. Distribuciones en las que esto ocurre se denominan *distribuciones de enlace canónico*. En la tabla 1 se presentan las principales distribuciones pertenecientes a la familia exponencial.

Distribución	$b(\theta)$	θ	ϕ	$V(\mu)$
Normal	$\theta^2/2$	μ	σ^{-2}	1
Poisson	e^{θ}	$\log(\mu)$	1	μ
Binomial	$\log(1+e^{\theta})$	$\log\{\mu/(1-\mu)\}$	n	$\mu(1-\mu)$
Gamma	$-\log(-\theta)$	$-1/\mu$	$1/(\mathrm{CV})^2$	μ^2
N. Inversa	$-\sqrt{-2\theta}$	$-1/2\mu^2$	ϕ	μ^3

Tabla 1: Principales distribuciones de la familia exponencial

1.1.2. Tipos de enlaces en GLM

Los enlaces canónicos para los modelos normal, binomial, Poisson, gamma e inversa gaussiana son respectivamente:

$$\mu=\eta, \quad \log\left\{\frac{\mu}{1-\mu}\right\}=\eta, \quad \log(\mu)=\eta, \quad \mu^{-1}=\eta, \quad \text{y} \quad \mu^{-2}=\eta$$

En la literatura estadística se encuentran también otro tipo de enlaces, a saber:

- **Enlace Probit:** Si μ es una proporción de sucesos de una binomial, el enlace probit se define como:
$$\Phi^{-1}(\mu)=\eta$$
en donde $\Phi(\cdot)$ es la función de distribución acumulada de una normal estándar.
- **Enlace Complemento log-log:** el modelo binomial con enlace log-log es definido por:
$$\mu=1-exp\{-exp(\eta)\})$$
o equivalentemente,
$$\log\{-\log(1-\mu)\}=\eta$$

- **Enlace Logístico:** utiliza como base la distribución logística y se define:

$$\mu = e^{\eta}/(1+e^{\eta})$$

- **Enlace de Box-Cox:** clase importante de enlaces por lo menos para observaciones positivas, se define:

$$\eta = (\mu^{\lambda} - 1)/\lambda$$

para $\eta \neq 0$ es $\eta = \log(\mu) \quad$ cuando $\quad \lambda \longrightarrow 0$.

- **Enlace de Aranda-Ordaz:** una transformación importante fué propuesta por Aranda-Ordaz para datos binarios (1981, [2]),

$$\eta = \log\left\{ \frac{(1-\mu)^{-\alpha} - 1}{\alpha} \right\}$$

en donde $0 < \mu < 1$ y α es una constante desconocida. Cuando $\alpha = 1$ se tiene el enlace logit $\eta = \log(\mu/(1-\mu))$. Cuando $\alpha \longrightarrow 0$ se tiene $\{(1-\mu)^{-\alpha} - 1\}/\alpha \longrightarrow \log(1-\mu)^{-1}$ de modo que $\eta = \log\{-\log(1-\mu)\}$, es decir el enlace complemento log-log.

1.1.3. Inferencia en un Modelo Lineal Generalizado (GLM)

La metodología estadística para la estimación de los parámetros en un GLM se basa en *máxima verosimilitud*. aquí se sigue estrictamente la metodología y notación utilizada por Paula (2004, [19], pp. 5). El logaritmo de verosimilitud con respuestas independientes se expresa:

$$L(\boldsymbol{\beta}; \mathbf{y}) = \sum_{i=1}^{n} \phi\{y_i\theta_i - b(\theta_i)\} + \sum_{i=1}^{n} c(y_i, \phi)$$

Cuando el parámetro canónico θ_i coincide con el predictor lineal, es decir cuando $\theta_i = \eta_i = \sum_{j=1}^{p} x_{ij}\beta_j$, la ecuación anterior se convierte en:

$$L(\boldsymbol{\beta}; \mathbf{y}) = \sum_{i=1}^{n} \phi\left\{ y_i \sum_{j=1}^{p} x_{ij}\beta_j - b\left(\sum_{j=1}^{p} x_{ij}\beta_j \right) \right\} + \sum_{i=1}^{n} c(y_i, \phi)$$

Al definir $S_j = \phi \sum_{i=1}^{n} Y_i x_{ij}$, la ecuación anterior se puede escribir:

$$L(\boldsymbol{\beta}; \mathbf{y}) = \sum_{\mathbf{j=1}}^{\mathbf{p}} \mathbf{s_j}\, \beta_{\mathbf{j}} - \phi \sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{b}\left(\sum_{\mathbf{j=1}}^{\mathbf{p}} \mathbf{x_{ij}} \beta_{\mathbf{j}} \right) + \sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{c}(\mathbf{y_i}, \phi)$$

Y por el teorema de factorización, la estadística $S = (S_1, S_2, \cdots, S_n)^T$ es una estadística suficiente y minimal para el vector de parámetros $\boldsymbol{\beta} = (\beta_1, \beta_2, \cdots, \beta_p)^T$. Los enlaces que corresponden a tales estadísticas son llamados *enlaces canónicos* y cumplen un papel importante en la teoría de los GLM, pues garantizan concavidad y resultados asintóticos importantes.

1.1.4. Función Score y Matriz de Información

Para obtener la función score del parámetro $\boldsymbol{\beta}$ se calcula la primera derivada de la función logaritmo de verosimilitud, Paula (2004, [19], pp. 15).

$$\begin{aligned}\partial\, L(\boldsymbol{\beta};\mathbf{y})/\partial\,\beta_{\mathbf{j}} &= \textstyle\sum_{i=1}^{n}\phi\Big\{y_i\frac{d\theta_i}{d\mu_i}\frac{d\mu_i}{d\eta_i}\frac{\partial\,\eta_i}{\partial\,\beta_j}-\frac{db(\theta_i)}{d\theta_i}\frac{d\theta_i}{d\mu_i}\frac{d\mu_i}{d\eta_i}\frac{\partial\,\eta_i}{\partial\,\beta_j}\Big\}\\ &= \textstyle\sum_{i=1}^{n}\phi\Big\{y_i\,V_i(d\mu_i/d\eta_i)x_{ij}-\mu_i\,V_i(d\mu_i/d\eta_i)x_{ij}\Big\}\\ &= \textstyle\sum_{i=1}^{n}\phi\Big\{\sqrt{\frac{\omega_i}{V_i}}(y_i-\mu_i)x_{ij}\Big\}\end{aligned}$$

en donde $\omega_i=(d\mu_i/d\eta_i)^2/V_i$ y la función score escrita en forma vectorial queda:

$$U(\boldsymbol{\beta})=\frac{\partial\,L(\boldsymbol{\beta};\mathbf{y})}{\partial\,\boldsymbol{\beta}}=\phi\,\mathbf{X}^T\mathbf{W}^{1/2}\,\mathbf{V}^{-1/2}(\mathbf{y}-\boldsymbol{\mu})$$

en donde $\mathbf{X}$ es una matriz de orden $n\times p$ de rango completo, $\mathbf{W}=\text{diag}\{\omega_{\mathbf{1}},\omega_{\mathbf{2}},\cdots,\omega_{\mathbf{n}}\}$ es una matriz de pesos, $\mathbf{V}=\text{diag}\{V_1,V_2,\cdots,V_n\}$, $\mathbf{y}=(y_1,y_2,\cdots,y_n)$ y $\boldsymbol{\mu}=(\mu_1,\mu_2,\cdots,\mu_n)$.

La matriz de información de Fisher se obtiene mediante el cálculo de las segundas derivadas parciales de la función logaritmo de verosimilitud ó de las primeras derivadas parciales de la función score:

$$\partial^2\,L(\boldsymbol{\beta};\mathbf{y})/\partial\beta_j\partial\beta_l=\phi\sum_{i=1}^{n}(y_i-\mu_i)\frac{d^2\theta_i}{d\mu_i^2}\left(\frac{d\mu_i}{d\eta_i}\right)^2 x_{ij}\,x_{il}+\phi\sum_{i=1}^{n}(y_i-\mu_i)\frac{d\theta_i}{d\mu_i}\frac{d^2\mu_i}{d\eta_i^2}x_{ij}\,x_{il}$$

$$-\phi\sum_{i=1}^{n}\frac{d\theta_i}{d\mu_i}\left(\frac{d\mu_i}{d\eta_i}\right)x_{ij}\,x_{il}$$

de donde su valor esperado es:

$$E(\partial^2\,L(\boldsymbol{\beta};\mathbf{y})/\partial\beta_j\partial\beta_l)=-\phi\sum_{i=1}^{n}\frac{d\theta_i}{d\mu_i}\left(\frac{d\mu_i}{d\eta_i}\right)x_{ij}\,x_{il}=-\phi\sum_{i=1}^{n}\frac{(d\mu_i/d\eta_i)^2}{V_i}x_{ij}\,x_{il}=-\phi\sum_{i=1}^{n}\omega_i x_{ij}x_{il}$$

y la matriz de información de Fisher en forma matricial se puede escribir:

$$\mathbf{K}(\boldsymbol{\beta})=E\Big\{-\frac{\partial^2 L(\boldsymbol{\beta};\mathbf{Y})}{\partial\boldsymbol{\beta}\,\partial\boldsymbol{\beta}^T}\Big\}=\phi\,\mathbf{X}^T\,\mathbf{W}\,\mathbf{X}$$

1.1.5. Estimación de $\boldsymbol{\beta}$.

Se utiliza el proceso iterativo de Newton-Raphson. y la estimación de máxima verosimilitud de $\boldsymbol{\beta}$ se define expandiendo la función *score* $U(\boldsymbol{\beta})$ alrededor de un valor inicial $\boldsymbol{\beta}^{(0)}$, es decir:

$$U(\boldsymbol{\beta})\approx U(\boldsymbol{\beta}^{(0)})+U'(\boldsymbol{\beta}^{(0)})(\boldsymbol{\beta}-\boldsymbol{\beta}^{(0)})$$

en donde $U'(\boldsymbol{\beta})$ representa la primera derivada de la función *score*. Al despejar $\boldsymbol{\beta}$ y repetir iterativamente el proceso se llega a:

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} + \{-U'(\boldsymbol{\beta}^{(m)})\}^{-1} U(\boldsymbol{\beta}^{(m)})$$

con $m = 0, 1, \cdots$ y como la matriz $U'(\boldsymbol{\beta}^{(m)})$ puede llegar a no ser definida positiva, ésta se cambia por la inversa de la matriz de información de Fisher, llegando al proceso iterativo:

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} + \mathbf{K}^{-1}(\boldsymbol{\beta}^{(m)})\} U(\boldsymbol{\beta}^{(m)})$$

con $m = 0, 1 \cdots$. Al realizar un poco de álgebra se llega a un proceso de mínimos cuadrados reponderados:

$$\boldsymbol{\beta}^{(m+1)} = (\mathbf{X}^T \, \mathbf{W}^{(m)} \, \mathbf{X})^{-1} \, \mathbf{X}^T \, \mathbf{W}^{(m)} \, \mathbf{z}^{(m)}$$

com $m = 0, 1, \cdots$ y $\mathbf{z} = \eta + \mathbf{W}^{-1/2} \, \mathbf{V}^{-1/2} \, (\mathbf{y} - \boldsymbol{\mu})$ es una variable dependiente modificada.

Sen y Singer (1993, [21], Cap. 7) afirman que bajo condiciones de regularidad , el estimador de máxima verosimilitud de $\hat{\boldsymbol{\beta}}$ es un estimador consistente y eficiente de $\boldsymbol{\beta}$. esto significa:

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \longrightarrow_d \; N_p(\mathbf{0}, \boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta})) \quad \text{cuando} \quad n \longrightarrow \infty$$

Además

$$\boldsymbol{\Sigma}(\boldsymbol{\beta}) = \lim_{n \to \infty} \frac{\mathbf{K}(\boldsymbol{\beta})}{n}$$

siendo $\boldsymbol{\Sigma}(\boldsymbol{\beta})$ una matriz definida positiva.

1.1.6. Función Desvío

La función *desvío* sin pérdida de generalidad representa la discrepancia entre la verosimilitud del modelo saturado con n parámetros y el modelo estimado con p parámetros. Si se representa el logaritmo de verosimilitud para el modelo estimado por

$$L(\boldsymbol{\mu}; \mathbf{y}) = \sum_{i=1}^{n} L(\mu_i; y_i)$$

se sabe que $\mu_i = g^{-1}(\eta_i)$ y $\eta_i = \mathbf{X}_i^T \boldsymbol{\beta}$. Para el modelo saturado ($p = n$) la función logaritmo de verosimilitud es:

$$L(\mathbf{y}; \mathbf{y}) = \sum_{i=1}^{n} L(y_i; y_i)$$

La calidad del ajuste de un GLM se evalúa a través de la función *desvío*:

$$D^*(\mathbf{y}; \hat{\boldsymbol{\mu}}) = \phi \, D(\mathbf{y}; \hat{\boldsymbol{\mu}}) = 2\{L(\mathbf{y}; \mathbf{y}) - L(\hat{\boldsymbol{\mu}}; \mathbf{y})\}$$

Interesantes propiedades de la función desvío pueden ser consultadas en Jorgensen (1987, [11]).

1.1.7. Estimación del parámetro de dispersión ϕ

Se puede demostrar que los parámetros $\boldsymbol{\beta}$ y ϕ son ortogonales, esto se evidencia en $E\left[\partial^2 L(\boldsymbol{\beta};\phi;\mathbf{y})/\partial\,\boldsymbol{\beta}\,\partial\,\phi\right] = 0$. Una consecuencia inmediata es la independencia asintótica entre sus estimaciones $\hat{\phi}$ y $\hat{\boldsymbol{\beta}}$. Al derivar el logaritmo de verosimilitud con respecto al parámetro ϕ e igualando a cero, se llega a la solución:

$$\sum_{i=1}^{n} c'(y_i,\hat{\phi}) = \frac{1}{2}\,D(\mathbf{y};\hat{\boldsymbol{\mu}}) - \sum_{i=1}^{n}\{y_i\hat{\theta_i}^{\,0} - b(\hat{\theta_i}^{\,0})\}$$

donde $D(\mathbf{y};\hat{\boldsymbol{\mu}})$ representa la función desvío sobre el modelo bajo investigación. La estimación de máxima verosimilitud para ϕ en los modelos normal e inversa gaussiana están dados por: $\hat{\phi} = n/D(\mathbf{y};\hat{\boldsymbol{\mu}})$. Para la distribución gamma es

$$2n\{\log(\hat{\phi}) - \psi(\hat{\phi})\} = D(\mathbf{y};\hat{\boldsymbol{\mu}})$$

en donde $\psi(\phi) = \Gamma'(\phi)/\Gamma(\phi)$ es una función digamma.

Un estimador preferido de ϕ está basado en la estadística de Pearson

$$\hat{\phi}^{-1} = \sum_{i=1}^{n}\{(y_i - \hat{\mu_i})/\hat{\mu_i}\}^2/(n-p)$$

La condición para este estimador es que $\boldsymbol{\beta}$ debe haber sido estimado consistentemente.

1.1.8. Hipótesis

En este documento sólo se hace referencia a hipótesis simples. Supóngase que se desea probar la hipótesis:

$$H_0:\ \boldsymbol{\beta} = \boldsymbol{\beta}^0 \quad \text{versus} \quad H_1:\ \boldsymbol{\beta} \neq \boldsymbol{\beta}^0$$

en donde $\boldsymbol{\beta}^0$ es un vector p-dimensional conocido y ϕ también se asume conocido. Las siguientes pruebas son utilizadas en GLMs:

- **Prueba de Razón de Verosimilitud:** se define

  $$\epsilon_{RV} = 2\{L(\hat{\boldsymbol{\beta}};\mathbf{y}) - L(\boldsymbol{\beta}^0;\mathbf{y})\}$$

  Esta estadística se define como la diferencia entre dos funciones desvío:

  $$\epsilon_{RV} = \phi\{D(\mathbf{y};\hat{\boldsymbol{\mu}}^0) - D(\mathbf{y};\hat{\boldsymbol{\mu}})\}$$

- **Prueba de Wald:** utiliza la estadística:

  $$\epsilon_W = \left[\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^0\right]^T \hat{\mathrm{Var}}(\hat{\boldsymbol{\beta}})\left[\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^0\right]$$

donde $\hat{\text{Var}}(\hat{\boldsymbol{\beta}})$ representa la matriz de varianzas-covarianzas asintótica de $\hat{\boldsymbol{\beta}}$ estimada en $\hat{\boldsymbol{\beta}}$. Para GLMs $\hat{\text{Var}}(\hat{\boldsymbol{\beta}}) = \mathbf{K}^{-1}(\hat{\boldsymbol{\beta}})$, así que la estadística se puede reescribir:

$$\epsilon_W = \left[\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^0\right]^T (\mathbf{X}^T \hat{\mathbf{W}} \mathbf{X}) \left[\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^0\right]$$

cuando el número de parámetros es igual a 1 la prueba de Wald es equivalente a una prueba t de Student.

- **Prueba Score:** conocida también como la prueba de Rao y se define cuando $U(\hat{\boldsymbol{\beta}}) = 0$. Para GLMs se define:

  $$\epsilon_{SR} = \phi^{-1}\mathbf{U}(\beta^{\mathbf{0}})^T(\mathbf{X^T}\hat{\mathbf{W}}_{\mathbf{0}}\mathbf{X})^{-1}\mathbf{U}(\beta^{\mathbf{0}})$$

  en donde $\hat{W}_0$ es estimada sobre H_0.

- **Prueba F:** esta prueba se define con base en la función desvío, esto es

  $$F = \frac{\{D(\mathbf{y};\hat{\boldsymbol{\mu}}^0) - D(\mathbf{y};\hat{\boldsymbol{\mu}})\}/p}{D(\mathbf{y};\hat{\boldsymbol{\mu}})/(n-p)} \sim F_{p,(n-p)} \quad \text{cuando} \quad \phi \longrightarrow \infty$$

Finalmente y sobre la hipótesis nula se tiene que ϵ_{RV}, ϵ_W y $\epsilon_{SR} \sim \chi^2_p$ y una región de confianza basada en la prueba de Wald para $\boldsymbol{\beta}$ y con $(1-\alpha)$ de confianza es:

$$(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T(\mathbf{X}^T \hat{\mathbf{W}} \mathbf{X})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \leq \phi^{-1}\chi^2_p(1-\alpha)$$

1.1.9. Residuos

Siguiendo los principios de Cook (1986, [6]), se conocen tres tipos de residuos, estos son:

- **Residuos simples:** se definen como la diferencia entre el valor observado de la variable respuesta y el valor estimado por el modelo. Tomando como referencia el modelo lineal se definen:

  $$\mathbf{r} = \mathbf{Y} - \hat{\mathbf{Y}}; \quad E(\mathbf{r}) = \mathbf{0}; \quad \text{y} \quad \text{Var}(\mathbf{r}) = (\mathbf{I} - \mathbf{H})\sigma^2$$

  en donde $\mathbf{I}$ es la matriz idéntica y $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ es llamada matriz Hat, tiene las propiedades de ser simétrica e idempotente, es decir $\mathbf{H}^T = \mathbf{H}$ y $\mathbf{H}^T\mathbf{H} = \mathbf{H}\mathbf{H}^T = \mathbf{H}$.

- **Residuos estandarizados:** son los mismos residuos simples estandarizados en media y varianza, para la i-ésima observación se notan y definen:

  $$t_i = \frac{r_i}{s(1-h_{ii})} \quad \text{con} \quad s^2 = \sum_{i=1}^{n} r_i^2/(n-p) \quad \text{para} \quad i = 1, 2, \cdots, n$$

- **Residuos estudentizados:** se definen como la diferencia entre el valor observado de la variable y el valor estimado, cuando la i-ésima observación ha sido eliminada, o sea $Y_i - \hat{Y_{(i)}}$, más exactamente se definen:

$$t_i^* = \frac{r_i}{s_{(i)}(1-h_{ii})^{1/2}}$$

$s_{(i)}^2$ es la varianza estimada sin la i-ésima observación.

En Rao (1973, [20], p.185) se establecen las siguientes relaciones:

$$s_{(i)}^2 = s^2\left(\frac{n-p-t_i^2}{n-p-1}\right) \quad \text{y} \quad t_i^*\left(\frac{n-p-1}{n-p-t_i^2}\right)$$

1.1.10. Influencia Local

Al multiplicar el vector de observaciones de un modelo por un vector $\boldsymbol{\delta} = (\delta_1, \delta_2, \cdots, \delta_n)^T$, donde δ_j es un tipo de perturbación, definida tal que $0 \leq \delta_j \leq 1$. Cuando $\delta_j = 1 \ \forall j$ no hay perturbación en el modelo. Si $\delta_j = 0$ significa que la j-ésima observación fué excluída. El estimador de mínimos cuadrados para el modelo lineal, cuando el mismo modelo ha sido perturbado se escribe:

$$\hat{\boldsymbol{\beta}}_\delta = (\mathbf{X}^T \, \boldsymbol{\Delta} \, \mathbf{X})^{-1} \mathbf{X}^T \, \boldsymbol{\Delta} \, \mathbf{y}$$

en donde $\boldsymbol{\Delta} = \text{diag}(\delta_1, \delta_2, \cdots, \delta_n)$ es la matriz diagonal de perturbaciones. La medida de influencia más conocida se basa en la región de confianza para el parámetro $\boldsymbol{\beta}$,

$$(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T (\mathbf{X}^T \mathbf{X})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \leq \ p s^2 F_{p,(n-p)}(\alpha)$$

Se mencionan a continuación tres medidas de influencia, las cuales permiten medir la sensibilidad de un modelo estimado:

- **Distancia de Cook:** Excluyendo la i-ésima observación del modelo se define:

$$D_{(i)} = \frac{(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})^T (X^T X)(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})}{p \, s^2} = t_i^2 \frac{h_{ii}}{(1-h_{ii})} \frac{1}{p}$$

- **DFFITS:** propuestos por Belsley, Kuh y Welsch (1980, [4]), se definen por:

$$DFFITS = \frac{|r_i|}{s_{(i)}(1-h_{ii})^{1/2}} \left\{\frac{h_{ii}}{(1-h_{ii})}\right\}^{1/2} = |t_i^*| \left\{\frac{h_{ii}}{(1-h_{ii})}\right\}^{1/2}$$

- **Alternativa a los DFFITS:** Atkinson (1985, [3]) propone una alternativa a los DFFITS que se definen para la i-ésima observación como:

$$C_i = \left\{\frac{(n-p)}{p} \frac{h_{ii}}{(1-h_{ii})}\right\}$$

1.1.11. Criterios de Información

La validez y calidad de un GLM se mide a través de la función *desvío* y los criterios de información.

- **Criterio de información de Akaike:** mide la calidad relativa del ajuste de un modelo estadístico a un conjunto de datos:

  $$\mathbf{AIC} = 2\,k - 2\log(\mathbf{L})$$

  en donde k es el número de parámetros del modelo y $\mathbf{L}$ es el máximo valor de la función de verosimilitud. Se prefiere el modelo con menor **AIC**.

- **Criterio de información de Bayes:** criterio para la selección de un modelo, entre un conjunto finito de modelos,

  $$\mathbf{BIC} = -2\ln(\mathbf{L}) + k\ln(n)$$

  al tener dos modelos, se prefiere el que menor **BIC** tenga.

- **Criterio de información de Hannan - Quinn:** es un criterio alternativo al **AIC**, se define:

  $$\mathbf{HQC} = n\log\left(\frac{\mathbf{RSS}}{n}\right) + 2\,k\log\log(n)$$

  en donde **RSS** es la reducción en sumas de cuadrados del error del modelo estimado.

1.2. Modelos Aditivos Generalizados (GAMs)

Siguiendo a Rigby et.al (2005, [**?**]), las técnicas de suavización se hicieron muy populares a finales de los años 80s. Hastie y Tibshirani (1990, [10]) introdujeron los GLM en esa franja. Wood (2000, [25]) contribuyó enormemente a su teoría y desarrollo.

Un modelo aditivo generalizado GAM es definido por:

$$\mathbf{Y} \overset{\text{ind}}{\sim} \xi\left(\boldsymbol{\mu}, \boldsymbol{\phi}\right)$$

$$\boldsymbol{\eta} = g(\boldsymbol{\mu}) = \mathbf{X}\,\boldsymbol{\beta} + \mathbf{s}_1(\mathbf{x}_1) + \mathbf{s}_2(\mathbf{x}_2) + \cdots, + \mathbf{s}_J(\mathbf{x}_J)$$

donde $\mathbf{s}_j$ es una función suave no paramétrica, aplicada a la covariable $\mathbf{x}_j$. La idea es llevar los datos a determinar la relación entre el predictor $\eta = g(\mu)$ y las variables explicativas, antes que forzarlos a una relación lineal ó polinomial. Los términos de suavizamiento introducen no linealidad en el modelo. La estimación de parámetros en este tipo de modelos es una versión *penalizada* de la estimación para GLMs.

1.2.1. Funciones suaves univariadas.

La mejor representación de una función suave univariada es a través del modelo:

$$y_i = f(x_i) + \epsilon_i \tag{1}$$

en donde y_i es una variable respuesta indexada para la $i-$ésima observación, x_i es una covariable. f es una función suave y los ϵ_i son variables aleatorias independientes e idénticamente distribuidas $N(0, \sigma^2)$. Además se supone que la variable independiente x_i está en el intervalo $[0, 1]$.

1.2.2. Regresión Spline.

Para estimar la función f, se requiere que esta función sea representada de tal manera que el modelo (1) siga siendo un modelo lineal en los parámetros. Esto puede ser posible escogiendo una *base*, definiendo a su vez un espacio de funciones de la cual f sea un elemento. Al escoger una de tales bases, cantidades de elementos de la misma pueden ser escogidas. Por ejemplo si $b_i(x)$ es la base i-ésima de algún espacio, f se puede representar por:

$$f(x) = \sum_{i=1}^{q} b_i(x)\, \beta_i \tag{2}$$

A manera de ejemplo si f es un polinomio de grado 4 y $b_1(x) = 1, b_2(x) = x, b_3(x) = x^2, \ldots, b_5(x) = x^4$. El modelo (1) queda:

$$y_i = \beta_1 + \beta_2\, x_i + \beta_3\, {x_i}^2 + \ldots + \beta_5\, {x_i}^4 + \epsilon_i$$

Es de aclarar que las funciones polinómicas son muy útiles en situaciones en las que el interés se enfoca en las propiedades de f en la vecindad de un punto especificado. pero cuando las preguntas se relacionan al dominio completo (especificamente el intervalo $[0, 1]$, las bases polinómicas pueden tener algunos problemas. Las bases *spline* se comportan muy bien en tales circunstancias, a causa de que tienen buenas aproximaciones teóricas. En este orden de ideas un *spline cúbico* es una curva hecha de secciones de polinomios cúbicos, juntados o conectados de manera que sean contínuos y que se puedan evaluar la primera y segunda derivada de los mismos. Los puntos en los cuales las secciones se juntan se denominan *knots* del spline. Típicamente estos *knots* deberían ser equiespaciados, a través del rango de los valores observados de x, ó cuantilas de la distribución de la variable x como si fuera una sola variable. Sean los nodos (knots) de localización notados por $\{{x_i}^* : i = 1, 2, \ldots, q-2\}$.
Existen muchas formas de escribir una base para splines cúbicos, sin embargo la aproximación más general se puede encontrar en los libros de Wahba (1990, [24]) y Gu (2002, [9]) (citados por Wood). Para estas bases $b_1(x) = 1, b_2(x) = x$ y $b_{i+2}(x) = R(x, {x_i}^*)$ para $i = 1, \ldots, q-2$.

$$R(x,z) = \left[(z-\frac{1}{2})^2 - \frac{1}{12}\right]\left[(x-\frac{1}{2})^2 - \frac{1}{12}\right]/4 - \left[(|x-z| - \frac{1}{2})^4 - \frac{1}{2}((|x-z| - \frac{1}{2})^2 + \frac{7}{240})\right]/24 \quad (3)$$

Al usar splines cúbicos para f, el modelo (1) en forma matricial se escribe $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, donde la $i-$ésima fila de la matriz $\mathbf{X}$ se puede escribir:

$$\mathbf{X}_i = [1, x_i, R(x_i, {x_1}^*), R(x_i, {x_2}^*), \ldots, R(x_i, {x_{q-2}}^*)]$$

En consecuencia el modelo puede ser estimado por mínimos cuadrados.

1.2.3. Grado de suavizamiento mediante regresión spline penalizada.

Para controlar el suavizamiento alterando la dimensión de las bases, consiste en mantener fija la dimensión de la base tanto como sea razonable, pero controlando la suavidad del modelo añadiendo un ondulamiento penalizado a la función objetivo de *mínimos cuadrados*. Se pretende minimizar

$$||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2 + \lambda \int_0^1 \left[f''(x)\right]^2 dx$$

donde la integral de la segunda derivada al cuadrado penaliza el modelo. El parámetro λ es el parámetro de suavizamiento, cuando $\lambda \longrightarrow \infty$ se lleva a una línea recta que estima f, cuando $\lambda = 0$ resulta una regresión spline no penalizada.

Como f es lineal en los parámetros, β_i, la función penalizada puede ser escrita como una forma cuadrática:

$$\int_0^1 \left[f''(x)\right]^2 = \boldsymbol{\beta}^T \mathbf{S} \boldsymbol{\beta}$$

donde $\mathbf{S}$ es una matriz de constantes conocida, un elemento de esta matriz se puede escribir $S_{i+2,j+2} = R({x_i}^* {x_j}^*)$ para $i, j = 1, \ldots, q-2$ mientras que las dos primeras filas y columnas de $\mathbf{S}$ son cero. De esta forma la regresión spline penalizada consiste en minimizar:

$$||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2 + \lambda \boldsymbol{\beta}^T \mathbf{S} \boldsymbol{\beta} \quad (4)$$

Para la estimación del parámetro λ se hace necesario estimar primero el parámetro $\boldsymbol{\beta}$. Así después de un ejercicio algebraico sencillo pero engorroso se llega al estimador de mínimos cuadrados penalizados de $\boldsymbol{\beta}$.

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{S})^{-1} \mathbf{X}^T \mathbf{y} \quad (5)$$

Para este tipo de modelos la matriz de influencia, ó matriz Hat, $\mathbf{A}$, para el modelo es:

$$\mathbf{A} = \mathbf{X}\left(\mathbf{X}^T \mathbf{X} + \lambda \mathbf{S}\right)^{-1} \mathbf{X}^T$$

Se debe recordar que $\hat{\boldsymbol{\mu}} = \mathbf{A}\,\mathbf{y}$ y que la fórmula (4) también se expresa:

$$\left\| \begin{bmatrix} \mathbf{y} \\ \mathbf{0} \end{bmatrix} - \begin{bmatrix} \mathbf{X} \\ \sqrt{\lambda}\mathbf{B} \end{bmatrix} \boldsymbol{\beta} \right\|^2 = ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2 + \lambda \boldsymbol{\beta}^T \mathbf{S} \boldsymbol{\beta}$$

donde $\mathbf{B}$ es cualquier raiz cuadrada de la matriz $\mathbf{S}$ tal que $\mathbf{B}^T\mathbf{B} = \mathbf{S}$. Esta matriz $\mathbf{B}$ puede ser obtenida por descomposición espectral ó usando la descomposición de Cholesky.

1.2.4. Validación cruzada para la obtención del parámetro de suavización λ.

Si λ es muy grande los datos serán sobre suavizados, si por el contrario λ es pequeño los datos serán poco o nada suavizados. En los dos casos la estimación $\hat{f}$ no será cerrada a la verdadera función f. Un criterio para escoger λ puede ser minimizar:

$$M = \frac{1}{n}\sum_{i=1}^{n}(\hat{f}_i - f_i)^2 \qquad \text{con} \qquad \hat{f}_i \equiv \hat{f}(x_i) \quad \text{y} \quad f_i \equiv f(x_i)$$

Sea $\hat{f}^{[-i]}$ el modelo ajustado para todos los datos excepto y_i, se define el puntaje de *validación cruzada* ordinario por:

$$\nu_o = \frac{1}{n}\sum_{i=1}^{n}(\hat{f}_i^{[-i]} - y_i)^2$$

En Wood(2000, [25], pp. 129) se demuestra que:

$$\mathrm{E}(\nu_o) = \frac{1}{n}\mathrm{E}\left(\sum_{i=1}^{n}(\hat{f}_i - f_i)^2\right) + \sigma^2$$

Se puede demostrar que para grandes muestras $\mathrm{E}(\nu_o) \approx \mathrm{E}(M) + \sigma^2$. Resulta un poco engorroso realizar una regresión spline cada vez que se elimina un dato, esto lleva a utilizar *validación cruzada generalizada* representada por:

$$\nu_g = \frac{n\sum_{i=1}.(y_i - \hat{f}_i)^2}{[tr(\mathbf{I} - \mathbf{A})]^2}$$

La *validación cruzada generalizada*cual tiene ventajas computacionales sobre la *validación cruzada ordinaria* en términos de invarianza, ver Wahba (1990, [24], pp. 53).

1.3. Modelos Aditivos de Localización, Escala y Forma (GAMLSS)

Los Modelos Aditivos Generalizados de Localización, Escala y Forma fueron propuestos inicialmente por Rigby y Stasinopoulos (2005). Se definen:

$$\begin{aligned}
&\mathbf{Y} \overset{\text{ind}}{\sim} \mathbf{D}\left(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\nu}, \boldsymbol{\tau}\right) \\
\boldsymbol{\eta}_1 &= g(\boldsymbol{\mu}) = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{s}_{11}(\mathbf{x}_{11}) + \mathbf{s}_{12}(\mathbf{x}_{12}) + \cdots, + \mathbf{s}_{1J_1}(\mathbf{x}_{1J_1}) \\
\boldsymbol{\eta}_2 &= g(\boldsymbol{\sigma}) = \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{s}_{21}(\mathbf{x}_{21}) + \mathbf{s}_{22}(\mathbf{x}_{22}) + \cdots, + \mathbf{s}_{2J_2}(\mathbf{x}_{2J_2}) \\
\boldsymbol{\eta}_3 &= g(\boldsymbol{\nu}) = \mathbf{X}_3\boldsymbol{\beta}_3 + \mathbf{s}_{31}(\mathbf{x}_{31}) + \mathbf{s}_{32}(\mathbf{x}_{32}) + \cdots, + \mathbf{s}_{3J_3}(\mathbf{x}_{3J_3}) \\
\boldsymbol{\eta}_4 &= g(\boldsymbol{\tau}) = \mathbf{X}_4\boldsymbol{\beta}_4 + \mathbf{s}_{41}(\mathbf{x}_{41}) + \mathbf{s}_{42}(\mathbf{x}_{42}) + \cdots, + \mathbf{s}_{4J_4}(\mathbf{x}_{4J_4})
\end{aligned}$$

Los parámetros $\boldsymbol{\nu}$ y $\boldsymbol{\tau}$ hacen referencia a la asimetría y curtosis de la distribución de los datos representados por la variable respuesta $\mathbf{Y}$. La estimación de parámetros en este tipo de modelos es similar a la de los GAMs.

1.3.1. Ventajas de los GAMLSS

- Crear una distribución *nueva* es relativamente *fácil*.
- Cualquier distribución puede ser *truncada* a derecha ó a izquierda.
- Una versión *censurada* de cualquier distribución puede ser creada.
- Cualquier distribución puede ser mezclada para crear mezclas finitas.
- Distribuciones contínuas *discretizadas* pueden ser creadas para modelar variables respuesta.
- Cualquier distribución contínua en el intervalo $(-\infty, \infty)$ puede ser transformada a una distribución en el intervalo $(0, \infty)$ ó el intervalo $(0, 1)$.

2. MARCO CONCEPTUAL

2.1. Medición de la inflación en Colombia.

El IPC en Colombia es calculado mensualmente por el Departamento Administrativo Nacional de Estadística (DANE) y está conformado por una canasta de bienes y servicios que estadísticamente representa el consumo de las familias colombianas y que resultan de una transacción de mercado. De esta medición se excluyen las trasferencias sociales en especie e impuestos entre otros, tal y como lo describe el DANE: "Las transferencias sociales en especie no se incluyen en la investigación, a menos que éstas se transformen y representen un gasto de consumo en el hogar como, por ejemplo, los servicios gratuitos ofrecidos por el gobierno que empiezan a ser cobrados dejando de lado su gratuidad. En segundo lugar, la adquisición del bien o servicio debe ser fruto de la decisión soberana del consumidor. Lo anterior determina que el alcance del IPC, está constituido por gastos de consumo final de los hogares, excluyendo erogaciones obligatorias como es el caso de los impuestos, y los seguros y contribuciones a la seguridad social. Los gastos de inversión y ahorro también son excluidos debido a que no son parte del gasto en consumo final", DANE (Ficha Técnica)[1].

Históricamente este IPC en Colombia se ha medido a partir de 1954 y hasta la fecha se han hecho 5 revisiones a la metodología, siendo la última realizada en 2008, fecha en la que se ajustó: la canasta de bienes, la ponderación del cálculo y la cobertura geográfica. Esta última revisión tiene como base la Encuesta de Ingresos y Gastos realizada entre 2006 y 2007. Al respecto es importante señalar que la medición de un índice de una canasta es diferente a un Índice de Costo de Vida (ICV). El IPC mide cuál ha sido el cambio que entre dos periodos ha tenido un conjunto o canasta de bienes, que para

[1] `https://www.dane.gov.co/files/investigaciones/fichas/DSO-IPC-FME-001-V7.pdf`

el Manual del índice de precios al consumidor se conoce como "Indice de Lowe". Esta metodología es diferente al Índice de Costo de Vida (ICV), que mide "el cambio en el costo mínimo de mantener un determinado nivel de vida" (2006, [18]).
Los resultados mensuales de este índice son publicados mensualmente en la página web del DANE en dos grupos, por Variaciones y por Índices tal y como se transcribe a continuación:

2.1.1. Variaciones.

- Variaciones Porcentuales IPC / 2002 – 2018.
- Variación mensual del IPC, por grupos de bienes y servicios / 216-2017. (Alimentos, Vivienda, Vestuario, Salud, Educación, Esparcimiento, Transporte, Comunicaciones y Otros gastos).
- Variación anual (12 meses) del IPC, por grupos de bienes y servicios /2016 – 2018.
- Variación mensual del IPC, según ciudades / 2016 – 2018 (Medellín, Barranquilla, Bogotá, Cartagena, Tunja, Manizales, Florencia, Popayán, Valledupar, Montería, Quibdó, Neiva, Riohacha, Santa Marta, Villavicencio, Pasto, Cúcuta, Armenia, Pereira, Bucaramanga, Sincelejo, Ibagué, Cali y San Andrés. (No se incluyen: Arauca, Guainía, Amazonas, Vaupes, Putumayo, Vichada, Guaviare y Casanare).
- Variación anual (12 meses del IPC, según ciudades / 2016 – 2018.
- PAAG. (Porcentaje de Ajuste Año Gravable).

2.1.2. Indices y Ponderaciones

Números índices y ponderaciones, por grupos de gastos nacional / 2018.

- Total ingresos / 2010 – 2018.
- Ingresos altos / 2010 – 2018.
- Ingresos medios / 2010 – 2018.
- Ingresos bajos / 2010 – 2018.

Cereales Y Productos de Panadería, Tubérculos y Plátanos, Hortalizas y Legumbres, Frutas, Carnes y Derivados de la Carne, Pescado y Otras de Mar, Lácteos, Grasas y Huevos, Alimentos varios, Comidas Fuera del Hogar, Gasto de Ocupación, Combustibles, Muebles del Hogar, Aparatos Domésticos,

Utensilios Domésticos, Roja del Hogar, Artículos para limpieza, Vestuario, Calzado, Servicios de vestuario, Servicios de salud, Bienes y artículos, Gastos de aseguramiento privado, Instrucción y enseñanza, Artículos escolares, Artículos culturales y otros artículos relacionados, Aparatos para diversión y esparcimiento, Servicios de diversión, Transporte personal, Transporte público, Comunicaciones, Bebidas alcohólicas, Artículos para el aseo y el cuidado persona, Artículos de joyería y otros personales y otros bienes y servicios.

- Número índices y ponderaciones, por clases de gasto nacional / 2018, (88 clases).
- Número índices y ponderaciones, por gastos básicos nacional / 2018, (181 clases).
- Índices, Series de Empalme / 2002 - 2018.

Los resultados de este indicador son utilizados por el gobierno y por entidades particulares para la toma de decisiones, entre otras como factor de ajuste para re-expresar cifras a precios constantes, lo anterior debido que la estabilidad actual de los precios en Colombia y la estructura de elaboración del IPC sirven como un buen predictor de precios debido a la rigidez que presenta su medición tal y como se menciona a continuación:

2.1.3. Rigidez de la inflación en Colombia.

Con mayor frecuencia se están realizando estudios sobre la rigidez de precios basados en la información del Índice de Precios al Productor (IPP) y el Índice de Precios al Consumidor (IPC), lo anterior con el propósito de evaluar el efecto que las políticas monetarias tienen sobre los precios y sobre la actividad económica. Una de las conclusiones de estos estudios ha sido que las empresas con mayor poder de mercado tienen mayor estabilidad en los precios en los bienes y productos que ofrecen, de otra parte, también los productos con mayor grado de elaboración tienen mayor rigidez y se conoce como "Snake effect en el cual los precios de los factores se transmiten lentamente al intermediario y al precio final del bien", Blanchard (2008, [5]).
Para Colombia se estimó la "Duración implícita" (meses) si cambió de cada grupo de bienes y servicios del IPC, encontrando que para los grupos vivienda, vestuario, educación y salud se detectaron rigideces mayores a 10 meses, tal y como se detalla en la tabla 2.
A su vez en este estudio se concluye que cuando la inflación en Colombia se acerque a su meta del largo plazo (3 %) la duración de las rachas se acercará entre 10 y 12 meses, entendiendo estas últimas como periodos de tiempo en los que bienes y/o servicios de la muestra permanecen inalterables en el tiempo, como es el caso de los bienes con control de precios. No sucede de esta forma con los alimentos perecederos, servicios públicos y transporte de combustible que presentan las mayores flexibilidades. De acuerdo con los resultados del estudio mencionado, esta tendencia se cerca a la tendencia mundial, pero con mayor flexibilidad para Colombia que para la zona euro y algunos países europeos.

Grupo	Duración implícita (rachas sin cambio de precios)
Alimentos	3.3
Vivienda	12.5
Transporte y comunicaciones	6.0
Otros gastos	7.1
Vestuario	10.3
Educación	16.0
Salud	13.1
Diversión y cultura	6.7

Tabla 2: Duración implícita para grupos de bienes y servicios.

Por lo tanto, esta conclusión que sugiere que la composición estructural del IPC modela su estabilidad. Esta condición muestra una relación inversa entre menor inflación y mayores rachas.

2.2. Característica general de las economías del Tolima y departamentos circunvecinos.

La economía de los departamentos alrededor del Tolima ha tenido al comienzo del siglo XX similitudes en el sector primario con énfasis en algunos productos como el Café para el eje cafetero y cultivos semestrales como el arroz para Tolima y Huila. Adicionalmente en todos, el sector terciario ha ganado participación en la conformación del PIB regional de estos departamentos. A continuación, se realiza una pequeña reseña de estas economías teniendo como referente estudios realizados por el Banco de la Republica en 2013.

2.2.1. Tolima, Huila y Cundinamarca.

Estos tres departamentos comparten áreas planas para el Tolima y Huila y la parte occidental del departamento de Cundinamarca. En estas áreas se comparten no solo el cultivo de productos transi-

torios sino hechos históricos y culturales similares en estos departamentos. De acuerdo con Campos Martínez, Álvaro y otros (2013)[2] la ganadería que el sector colonial fue importante para el desarrollo de la época ha perdido importancia y ha sido desplazada en importancia por la explotación petrolera que ha pasado a ser un reglón importante de la economía de estos departamentos.
Así mismo la región restante del departamento de Cundinamarca en el sector primario es muy importante con la ganadería especialmente con los hatos lecheros ya que son los más tecnificados del país. Un hecho importante a resaltar de esta región de Cundinamarca es que es el departamento más poblado del país y con la mayor contribución al PIB.[2]

2.2.2. Caldas, Quindío y Risaralda.

Esta región conocida como el eje cafetero tuvo en buen parte del siglo XX dependencia económica del Café. De acuerdo con Vallecilla (2011, [22]), hacia 1930 el 75 % de los cultivos agrícolas de ésta región era el Café, situación que cambió a finales del siglo XX cuando la Organización Internacional del Café OIC, acabó con su sistema de cuotas. Actualmente en el eje cafetero, el sector terciario tiene el 50 % de la contribución de la región en el PIB en actividades de servicios y bancaria principalmente. Adicionalmente el sector secundario con las actividades industriales y de alimentos contribuye con el 25 %.

2.2.3. Cauca y Valle del Cauca.

Estos dos departamentos ubicados al suroccidente del país, se encuentran en una región estratégica por estar en la cuenca del Pacífico que tiene un alto dinamismo con un mercado de 2.700 millones de habitantes con la generación del 55 % del PIB mundial. Esta región con sus condiciones de climáticas y de fauna, han contribuido a su crecimiento económico especialmente en el sector agrícola. No obstante lo anterior las zonas cercanas al Pacífico presentan altos niveles de pobreza y desigualdad que de acuerdo con Viloria (2007, [23]) son atribuibles "a su aislamiento geográfico, la falta de medios de transporte, escasez de tierras aptas en zona costera, la excesiva precipitación pluvial, elevada humedad, temperatura y proliferación de enfermedades endémicas".
En esta región se encuentra el puerto de Buenaventura el cual entre 1940 y 1950 se consolidó como el puerto más importante del país concentrado más del 60 % de las exportaciones del país y el 84 % de las ventas de Café. En este mismo periodo surgieron los ingenios azucareros que dieron vitalidad a la economía de la región. Al terminar el siglo se expidió la Ley Páez que otorgaba incentivos tributarios a las empresas constituidas en el Cauca y Huila, hecho que permitió dinamizar no solo la actividad

[2]Campos, A.A.; Martínez, A.A; Ramírez H. y Quintero,P.E. *Composición de la economía de la región centro de Colombia.*

industrial, sino la diversificación de la economía.

2.3. Comportamiento de la inflación en el periodo de estudio.

Durante el periodo de estudio el IPC registrado en los departamentos del presente trabajo ha presentado un comportamiento estable salvo el periodo 2016 cuando presentó un promedio de inflación para estos departamentos de 8.94 %, cifra similar a la reportada por el DANE para el país con 8.6 %. En este periodo la causa principal de la inflación fueron los alimentos 14.28 % en todo el país, lo anterior originado por el fenómeno del niño que afectó este grupo de la canasta familiar en el primer semestre de 2016 y el paro camionero que afectó este grupo por 45 días hasta el 19 de julio de 2016, siendo el más largo de la historia colombiana.

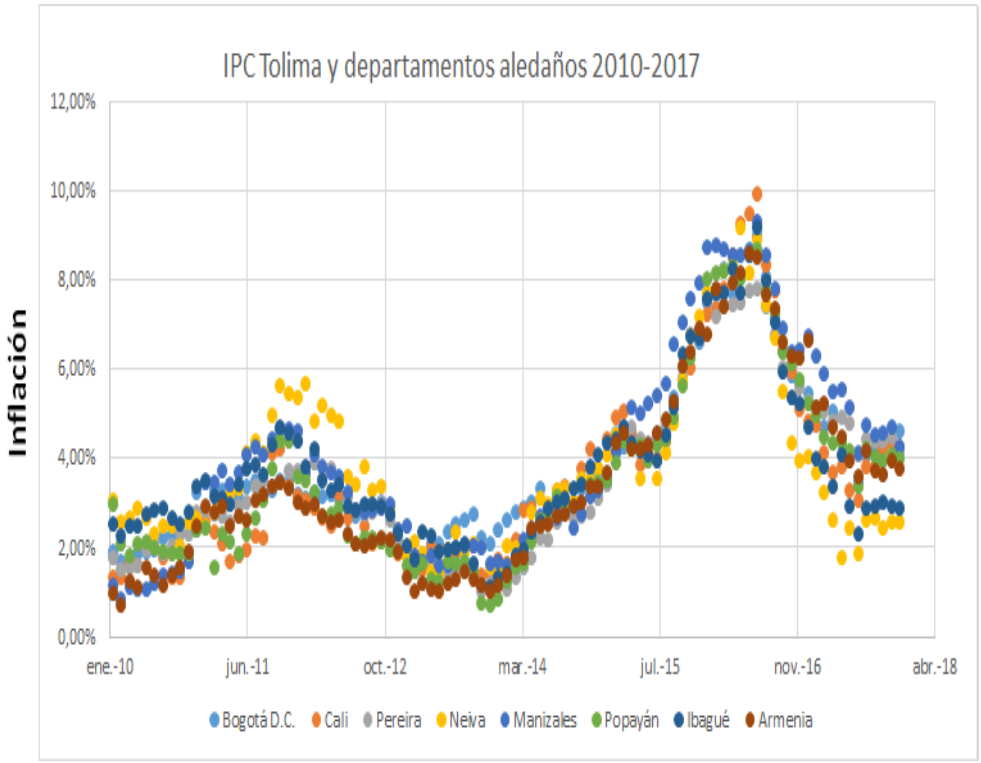

Figura 1: Comparaciones del IPC Tolima y departamentos circunvecinos 2010 al 2017.

3. ANÁLISIS DE RESULTADOS

3.1. Modelos de comparación Tolima vs Departamentos circunvecinos.

El Departamento Administrativo Nacional de Estadística (DANE) tiene los registros del Indice de Precios al Consumidor (IPC) en cada una de las ciudades capitales de los 23 Departamentos de Colombia, como se mencionó en el capítulo anterior, éste índice puede variar de región a región.
El Departamento del Tolima gracias a estar en el centro occidente del país, se considera un corredor importante que une la capital del país con el puerto de Buenaventura, en donde se mueve un importante porcentaje de la economía del país.
Debido a que el índice de precios es una cantidad pequeña, la metodología estadística más apropiada en este tipo de datos es la regresión *spline*, la cual se explica en los referentes teóricos. Como el IPC se toma en las ciudades capitales, en el presente trabajo se habla en forma indistinta del departamento ó de su capital. En las Tablas 3,4,5,6,7,8, y 9 se relacionan el IPC del Tolima, con cada uno de los departamento circunvecinos.

Fuente de Variación	Estimate	Std. Error	t value	Pr(>|t|)	Signif.
(Intercept)	0.0284	0.0035	8.13	0.0000	***
bSpline(x, df = 5)1	-0.0176	0.0062	-2.85	0.0054	**
bSpline(x, df = 5)2	0.0080	0.0036	2.22	0.0289	*
bSpline(x, df = 5)3	0.0057	0.0063	0.89	0.3748	
bSpline(x, df = 5)4	0.0560	0.0071	7.87	0.0000	***
bSpline(x, df = 5)5	0.0588	0.0058	10.11	0.0000	***

Tabla 3: Análisis regresión spline Ibagué vs Bogotá

En la tabla 3 se evidencia que el IPC para el Tolima (y) se puede explicar por medio del IPC Cundinamarca (x), mediante la ecuación:

$$\hat{y} = 0{,}0284 - 0{,}0176\,x + 0{,}056\,x^4 + 0{,}0588\,x^5$$

Con lo anterior se puede concluír que el intersecto y los polinomios de grados 4 y 5 aplicados al IPC de Cundinamarca (Bogotá), explican el IPC del departamento del Tolima, representado por su capital Ibagué, también hay una sección importante del IPC del Tolima, explicado en menor grado por

un polinomio de grado 1, el polinomio de grado 2 es aún menos significante, mientras que el polinomio de grado 3 no es significativo.

Fuente de Variación	Estimate	Std. Error	t value	Pr(>|t|)	Signif.
(Intercept)	0.023080	0.003982	5.796	9.91e-08	***
bSpline(x, df = 5)1	-0.003219	0.005853	-0.550	0.5837	
bSpline(x, df = 5)2	0.009395	0.004148	2.265	0.0259	*
bSpline(x, df = 5)3	0.014487	0.007000	2.070	0.0413	*
bSpline(x, df = 5)4	0.069549	0.007231	9.618	1.78e-15	***
bSpline(x, df = 5)5	0.061920	0.006189	10.005	2.79e-16	***

Tabla 4: Análisis regresión spline Ibagué vs Cali

La tabla 4 permite extraer el modelo de regresión *spline* estimado:

$$\hat{y} = 0{,}023080 + 0{,}069549\, x^4 + 0{,}061920\, x^5$$

en este caso son el intersecto y los polinomios de grado 4 y 5 los que mejor explican el IPC del Tolima, a través del IPC Valle del Cauca.

Fuente de Variación	Estimate	Std. Error	t value	Pr(>|t|)	Signif.
(Intercept)	0.0130	0.0029	4.48	0.0000	***
bSpline(x2, df = 5)1	0.0071	0.0049	1.44	0.1533	
bSpline(x2, df = 5)2	0.0256	0.0032	8.04	0.0000	***
bSpline(x2, df = 5)3	0.0174	0.0050	3.50	0.0007	***
bSpline(x2, df = 5)4	0.0563	0.0055	10.33	0.0000	***
bSpline(x2, df = 5)5	0.0726	0.0040	18.11	0.0000	***

Tabla 5: Análisis regresión spline Ibagué vs Pereira

La tabla 5 relaciona el modelo de regresión *spline* estimado:

$$\hat{y} = 0{,}0130 + 0{,}0256\, x^2 + 0{,}0174\, x^3 + 0{,}0563\, x^4 + 0{,}0726\, x^5$$

en este caso el polinomio de grado 1 no es significativo.

Fuente de Variación	Estimate	Std. Error	t value	Pr(>|t|)	Signif.
(Intercept)	0.0326	0.0034	9.57	0.0000	***
bSpline(x3, df = 5)1	-0.0231	0.0058	-3.98	0.0001	***
bSpline(x3, df = 5)2	0.0044	0.0037	1.21	0.2280	
bSpline(x3, df = 5)3	-0.0002	0.0059	-0.03	0.9761	
bSpline(x3, df = 5)4	0.0331	0.0060	5.55	0.0000	***
bSpline(x3, df = 5)5	0.0331	0.0051	11.25	0.0000	***

Tabla 6: Análisis regresión spline Ibagué vs Manizales

La tabla 6 permite evidenciar el modelo de regresión *spline*:

$$\hat{y} = 0{,}0326 - 0{,}0231\,x + 0{,}0331\,x^4 + 0{,}0331\,x^5$$

con lo que el intersecto y los polinomios de grados 1, 4 y 5 son los más significativos, que explican el IPC del Tolima a través del IPC Caldas.

Fuente de Variación	Estimate	Std. Error	t value	Pr(>|t|)	Signif.
(Intercept)	0.0115	0.0032	3.58	0.0006	***
bSpline(x4, df = 5)1	0.0048	0.0049	0.98	0.3305	
bSpline(x4, df = 5)2	0.0231	0.0037	6.26	0.0000	***
bSpline(x4, df = 5)3	0.0240	0.0052	4.58	0.0000	***
bSpline(x4, df = 5)4	0.0633	0.0060	10.62	0.0000	***
bSpline(x4, df = 5)5	0.0739	0.0044	16.84	0.0000	***

Tabla 7: Análisis regresión spline Ibagué vs Popayán

En la tabla 7 se visualiza el modelo estimado:

$$\hat{y} = 0{,}0115 + 0{,}0231\,x^2 + 0{,}0240\,x^3 + 0{,}0633\,x^4 + 0{,}0739\,x^5$$

y se evidencia que el único polinomio no significativo es el de grado 1.

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	0.0250	0.0042	5.96	0.0000	***
bSpline(x5, df = 5)1	-0.0113	0.0070	-1.62	0.1086	
bSpline(x5, df = 5)2	0.0117	0.0041	2.84	0.0056	**
bSpline(x5, df = 5)3	0.0074	0.0066	1.12	0.2661	
bSpline(x5, df = 5)4	0.0434	0.0063	6.92	0.0000	***
bSpline(x5, df = 5)5	0.0642	0.0056	11.45	0.0000	***

Tabla 8: Análisis regresión spline Ibagué vs Armenia

De acuerdo a la tabla 8, el modelo estimado de regresión *spline* es:

$$\hat{y} = 0{,}0250 + 0{,}0117\,x^2 + 0{,}0434\,x^4 + 0{,}0642\,x^5$$

lo que indica que intersecto y polinomios de grado 2, 4 y 5 explican el IPC Tolima, a través del IPC Quindío. A pesar de la cercanía con Ibagué no es un buen modelo.

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	0.0149	0.0050	2.98	0.0037	**
bSpline(x6, df = 5)1	0.0027	0.0075	0.36	0.7201	
bSpline(x6, df = 5)2	0.0157	0.0050	3.15	0.0022	**
bSpline(x6, df = 5)3	0.0279	0.0074	3.78	0.0003	***
bSpline(x6, df = 5)4	0.0644	0.0077	8.31	0.0000	***
bSpline(x6, df = 5)5	0.0717	0.0068	10.56	0.0000	***

Tabla 9: Análisis regresión spline Ibagué vs Neiva

La tabla 9 permite escribir el modelo de regresión estimado:

$$\hat{y} = 0{,}0149 + 0{,}0157\,x^2 + 0{,}0279\,x^3 + 0{,}0644\,x^4 + 0{,}0717\,x^5$$

y se evidencia que sólo el polinomio de grado 1 no es significativo, los demás polinomios si lo son.

3.2. Análisis de parámetros, modelos estimados.

En la tabla 10 se resumen los parámetros asociados a cada uno de los modelos anteriores, en los cuales se intenta expicar el IPC del Tolima con los IPC´s de los departamentos circunvecinos.

Modelo	Smoot.Spar	λ	G.L.	R^2	RSS.Pen	$\hat{\sigma}_e^2$	GCV
Ibagué vs Bogotá	0.8133149	0.0002423	9.21	0.8906	0.002479	0.006043	3.4173e-05
Ibagué vs Cali	1.000663	0.004078	5.03	0.8888	0.003164	0.006093	3.8030e-05
Ibagué vs Pereira	0.659603	1.8479e-0.5	16.24	0.9977	0.0013	0.0049	2.2031e-0.5
Ibagué vs M/zales	0.8185	0.00047	8.21	0.9122	0.0023	0.0054	3.0100e-05
Ibagué vs Popayán	0.9377	0.0021	5.85	0.9257	0.0020	0.0049	2.5609e-05
Ibagué vs Armenia	0.8706	0.0011	6.64	0.9033	0.0027	0.0056	3.4317e-05
Ibagué vs Neiva	0.8790	0.0006	7.63	0.8890	0.0027	0.0060	3.7343e-05

Tabla 10: Parámetros estimados regresión spline Tolima vs Departamentos circunvecinos.

Los parámetros expuestos en la tabla 10 son en su orden:

- Grado de suavizamiento (Smooth.Spar).
- Parámetro de rugosidad (λ).
- Grados de libertad aproximados (G.L.).
- Coeficiente de determinación (R^2).
- Reducción en sumas de cuadrados penalizada (RSS.Pen).
- Varianza estimada de los errores ($\hat{\sigma}_e^2$).
- Coeficiente de validación cruzada generalizado (GCV).

Al observar los anteriores resultados se puede evidenciar un excelente comportamiento en orden descendente de los modelos Ibagué vs Pereira, seguido por Ibagué vs Popayán y luego Ibagué vs Manizales, tienen los coeficientes de determinación más grandes y la varianza estimada de los errores es más pequeña que la de los demás modelos, así como la reducción en sumas de cuadrados penalizada. Esto es un indicio de que las distancias cortas no relacionan la economía entre regiones.

3.3. Análisis gráfico modelos comparativos.

Los siguientes gráficos permiten visualizar el comportamiento de los datos y a su vez la validación gráfica y estudio de influencia de los mismos.

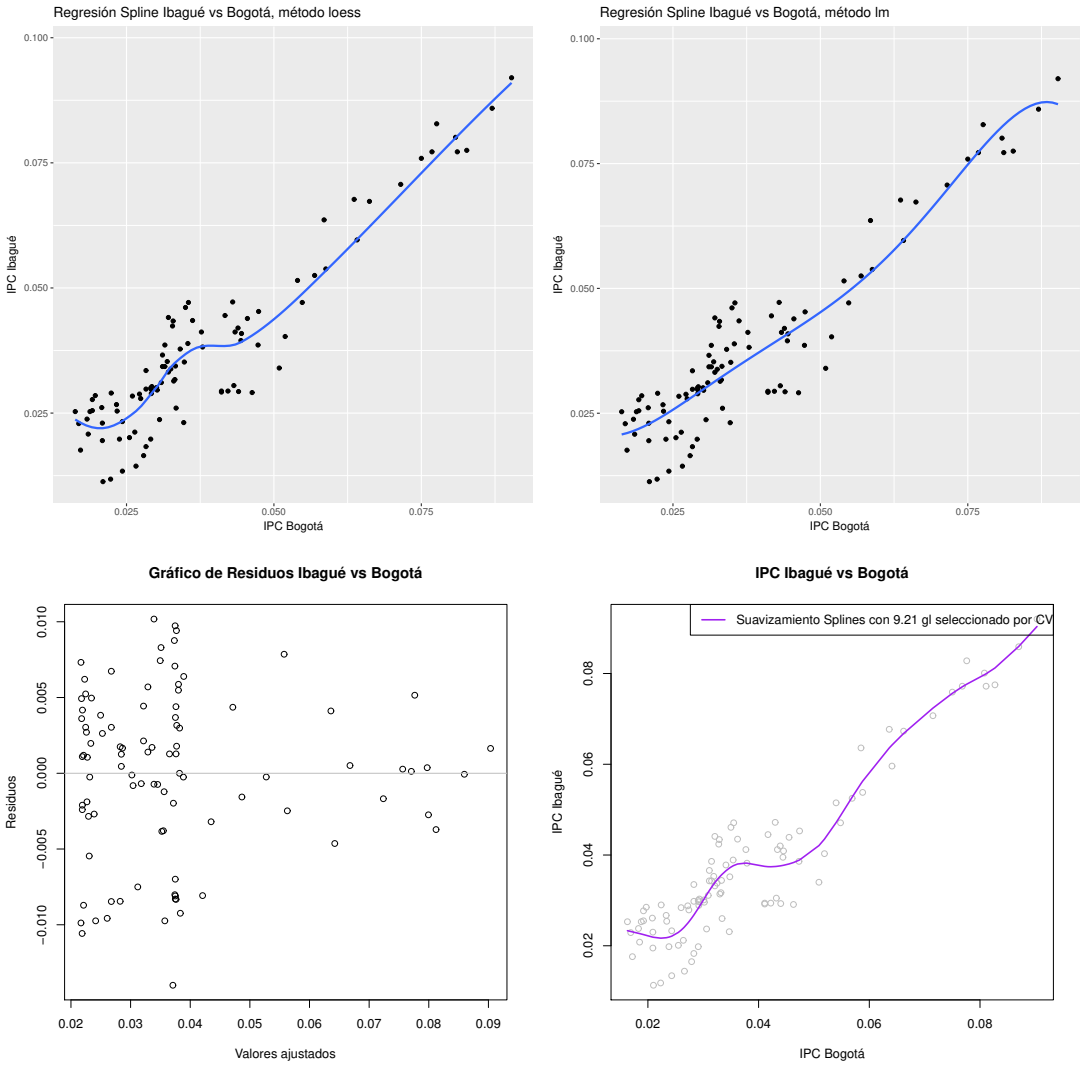

Figura 2: Modelo Ibagué vs Bogotá

En la figura 2.1 se observa la curva de ajuste del la regresión *spline* usando el método no paramétrico *loess*, en la figura 2.2 se observa también una curva de ajuste pero utilizando mínimos cuadrados, se

observa que quedan muchos puntos por fuera de la curva de ajuste. En la figura 2.3 se comparan los residuos de Ascombe y Tukey versus los valores ajustados del modelo, finalmente en la figura 2.4 se observa la curva final ajustada para el conjunto de datos, sugiriendo una función polinomial de grado 9.21, que explicaría mejor la relación entre estos dos IPC.

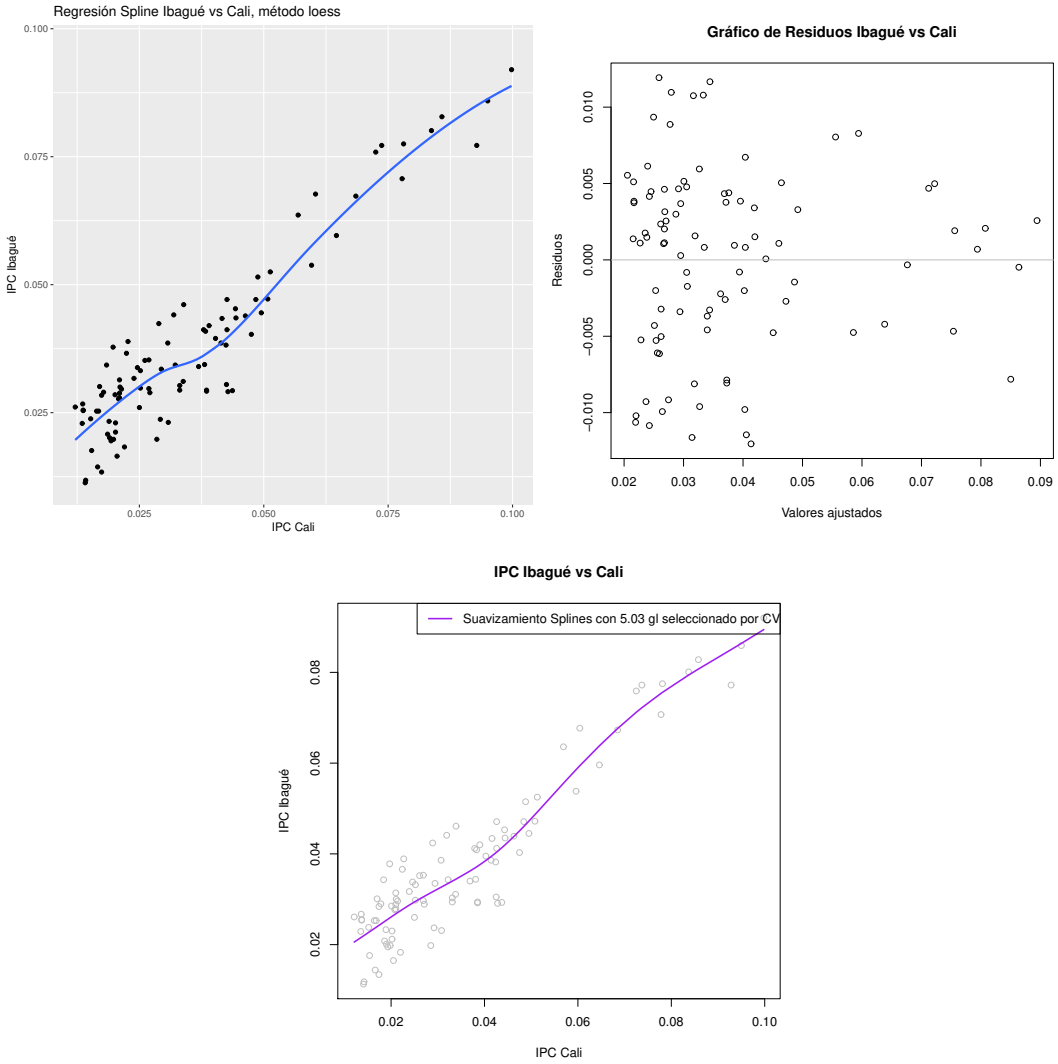

Figura 3: Modelo Ibagué vs Cali

En la figura 3.2 se observa que los residuos están en el intervalo [-0.010 , 0.010], esto es un rango muy pequeño e indica que el ajuste del modelo de comparación Ibagué vs Cali es bueno y que los valores

atípicos son muy pocos. Por su parte la figura 3.3 indica que el coeficiente de validación cruzada estuvo basado en una regresión spline con 5.3 grados de libertad aproximadamente, esto implica en forma subyacente que un polinomio de grado 5 ajusta muy bien el modelo asociado conjunto de datos que relaciona los IPC's Tolima vs Valle del Cauca.

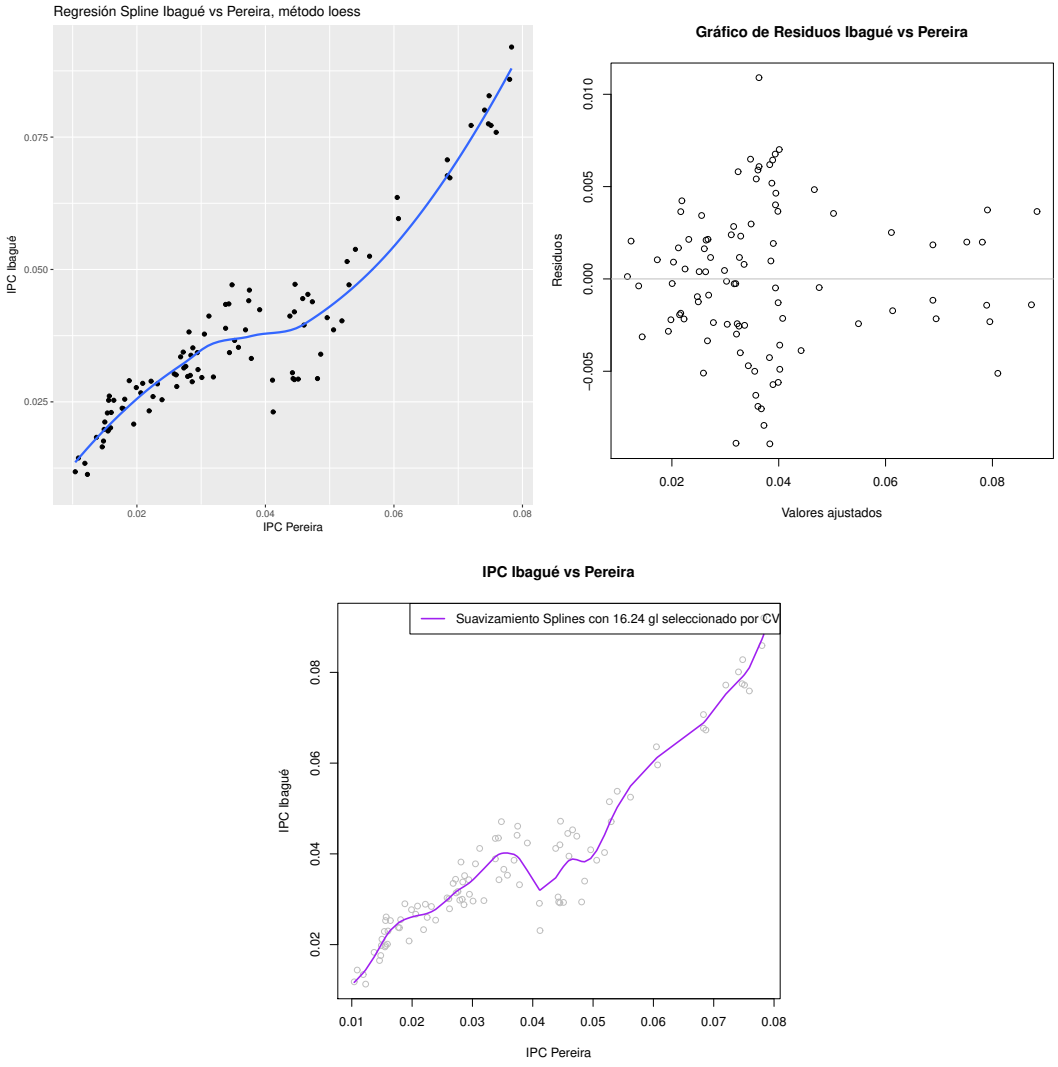

Figura 4: Modelo Ibagué vs Pereira

En la figura 4.3 se observa la función de ajuste *spline*, la cual posee mayor rugosidad, esto permite ver que el modelo estimado recoge mayor información de los datos, lo cual a su vez minimiza la variabili-

dad de los mismos. También se puede evidenciar en la figura 4.2 que los residuos estimados en su gran mayoría se encuentran en el intervalo [-0.005 , 0.005], no obstante algunos residuos estimados negativos se agrupan en los intervalos [-0.010 , -0.005] y algunos residuos positivos se agrupan en el intervalo [0.05 , 0.010], sin embargo el total de estos valores no supera el número 10. Estos valores corresponderían a observaciones atípicas ó valores influenciales de los datos, los cuales si se les observara a través de un modelo clásico se deberían intervenir. No obstante cuando se analizan los parámetros en la sección anterior se visualiza claramente que el modelo Ibagué vs Pereira es el mejor de todos los modelos analizados en el conjunto total de datos relacionados al IPC.

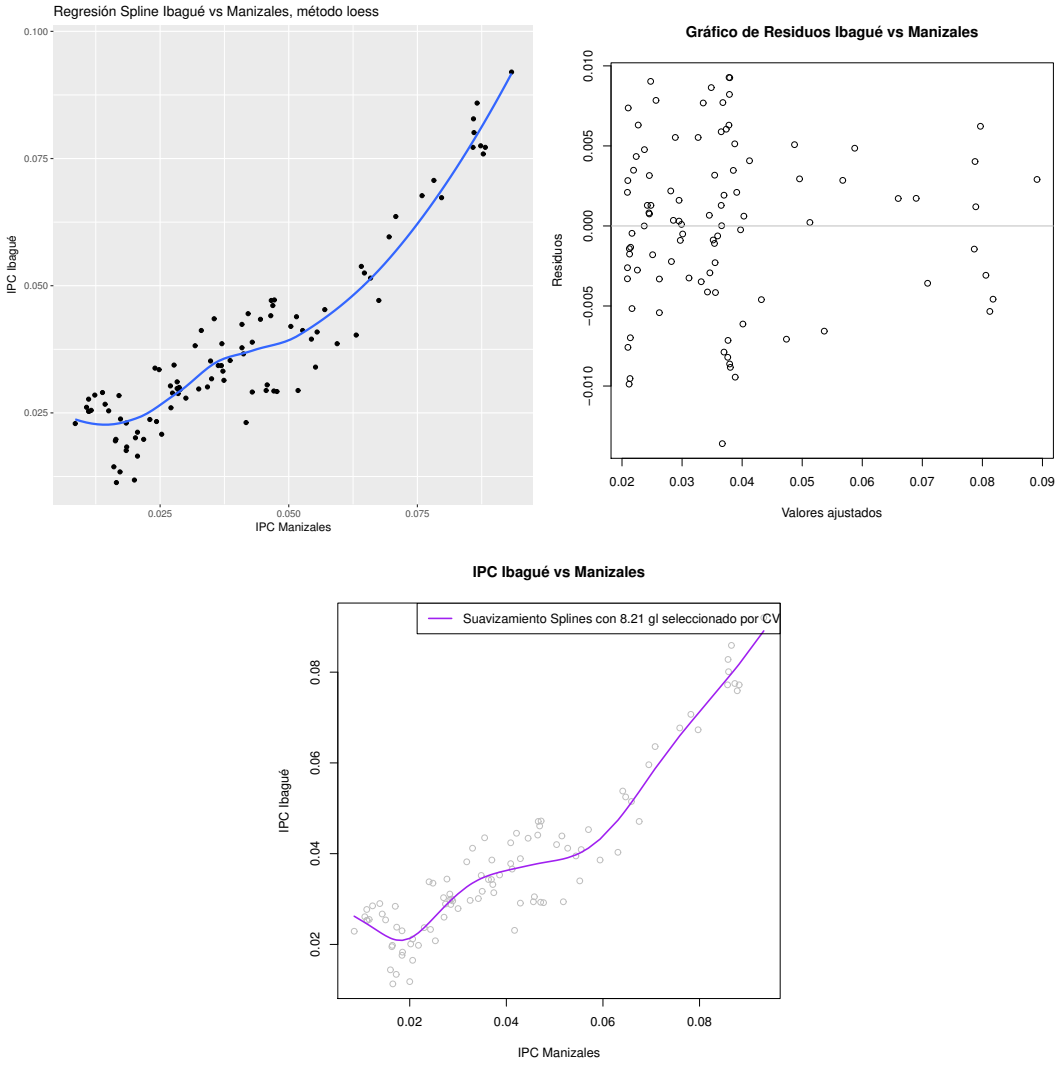

Figura 5: Modelo Ibagué vs Manizales

En las figuras 5.3 y 5.2 se pueden observar características similares pero no parecidas, la función *spline* estimada es menos rugosa que la del modelo inmediatamente anterior. Los residuos estimados nuevamente se agrupan en el intervalo [-0.010 , 0.010], siendo un poco mayores que el modelo que lo antecede. En forma muy general se puede apreciar que el modelo Ibagué vs Manizales es un buen modelos, no obstante no se puede considerar el mejor de todos. Con base en las características expuestas por los parámetros en la sección anterior, se puede concluír que éste modelo si es uno de los mejores, y que permite explicar en buena forma el IPC del Tolima, a través del IPC del Departamento de Caldas.

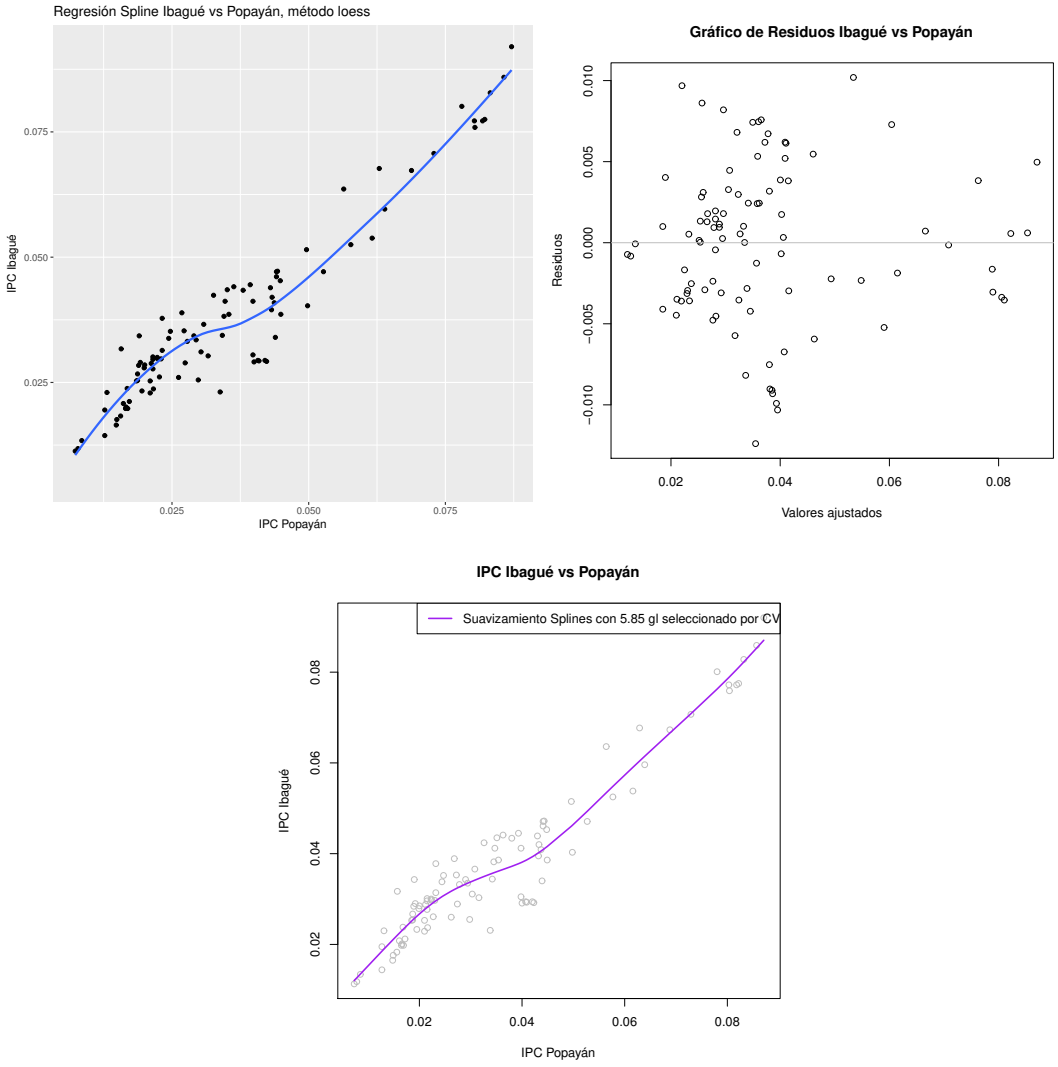

Figura 6: Modelo Ibagué vs Popayán

En las figuras 6.2 y 6.3 bien se podrían extraer conclusiones parecidas a las del modelo anterior. Los residuos estimados una vez más se agrupan en el intervalo [-0.010 , 0.010]. La rugosidad de la función *spline* estimada es muy similar a la del modelo anterior, no obstante la estimación de los parámetros en la sección anterior muestran que el coeficiente de determinación en el modelo anterior es 0.9122, en el presente modelo es de 0.9257, lo cual significa que este modelo es ligeramente superior al de Ibagué vs Manizales y por tanto pasa a formar parte de los mejores modelos.

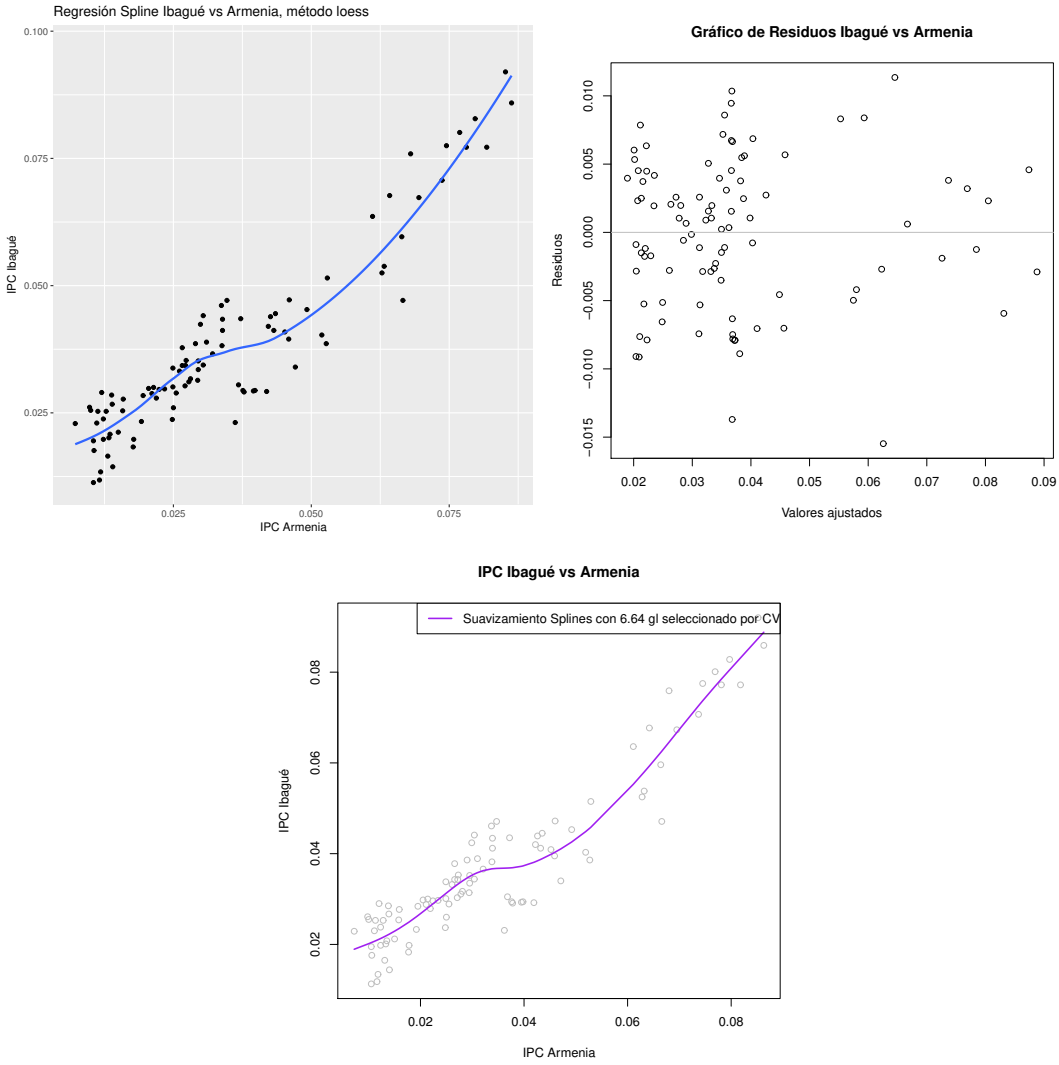

Figura 7: Modelo Ibagué vs Armenia

De acuerdo a las figuras 7.2 y 7.3, se evidencia que el gráfico de residuales estimados tiende a tener mayor dispersión, en forma positiva por encima de 0.010 y en forma negativa por debajo de -0.010. La función *spline* estimada por su parte, si tiene cierto parecido a la de los dos modelos anteriores, tiene menos rugosidad que en los dos modelos anteriores y esto se aprecia en los datos que no son explicados por el modelo en la parte baja de la gráfica 7.3. Al examinar los parámetros estimados de este modelo en la sección anterior, se aprecia que el coeficiente de determinación es de 0.9033, lo que a su vez implica que el modelo Ibagué vs Armenia es superado por los modelos Ibagué vs Pereira, Ibagué vs Popayán e Ibagué vs Pereira que hasta el momento vendría siendo el mejor de los modelos comparativos estimados. En este orden de ideas estaríamos pasando a un segundo umbral de buenos modelos.

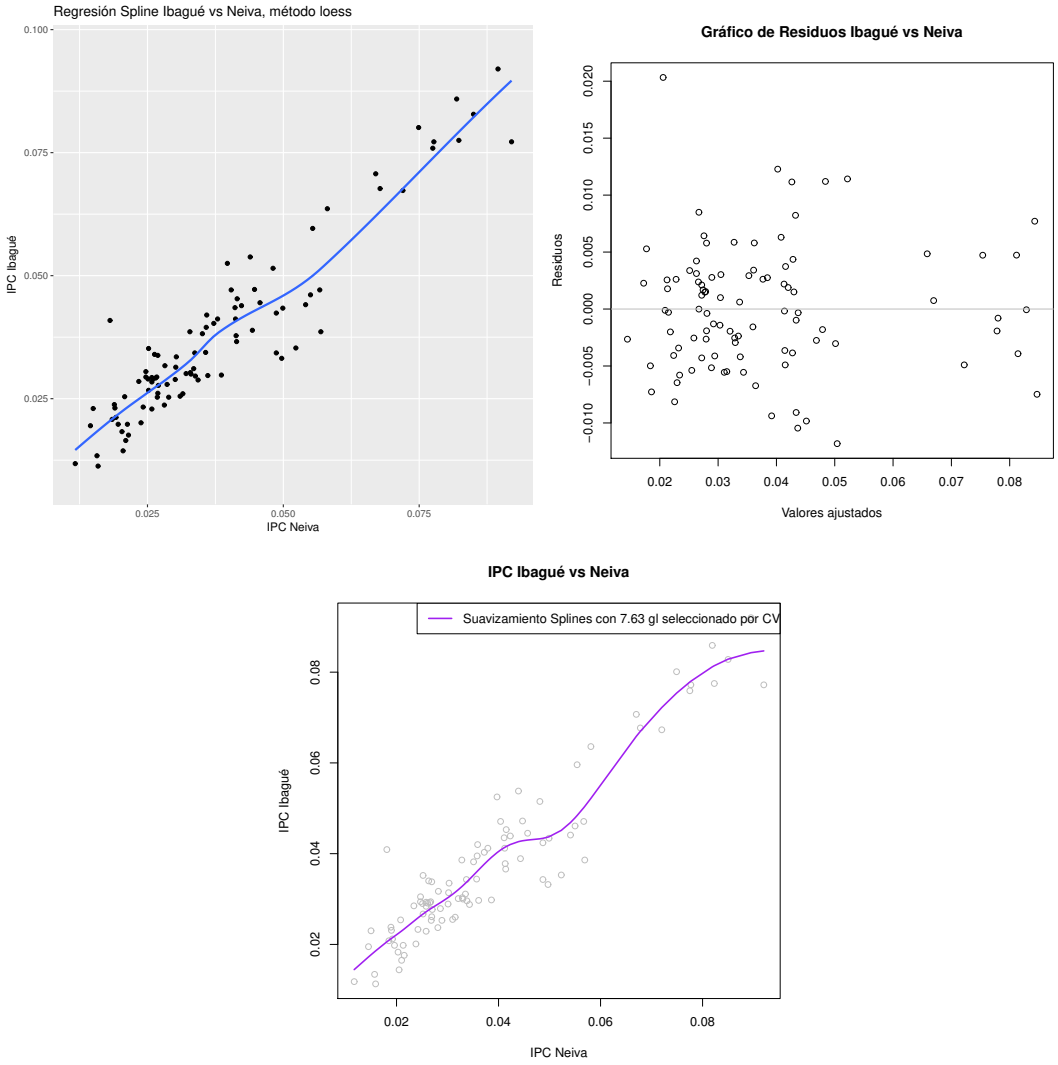

Figura 8: Modelo Ibagué vs Neiva

Por último en las figuras 8.2 y 8.3 se evidencian unas características que parecerían similares a las anteriores, sinembargo los residuos estimados positivos superan el valor 0.015 y los negativos se encuentran por debajo de -0.010. Al realizar un análisis del coeficiente de determinación, éste pasaría a ser el cuarto en orden descendente, quedando este modelo en el grupo de los segundos mejores por debajo del inmediatamente anterior.

4. Conclusiones

Con base en los análisis propuestos en el presente trabajo, se pueden escribir entre otras las siguientes conclusiones:

- El mejor modelo obtenido para explicar el Indice de Precios al Consumidor en el Tolima es el que relaciona Ibagué vs Pereira. Esto se debe quizás a la transferencia permanente de bienes de consumo: ropa, alimentos, bebidas, productos no perecederos entre estas dos ciudades por vía terrestre.
- El segundo mejor modelo para explicar el IPC en el Tolima, relaciona a Ibagué vs Popayán, esto se debe quizás a que una gran cantidad de productos precederos tales como la papa y algunos cultivos de tierra fría provienen de esa región.
- El tercer modelo relaciona el IPC del Tolima con Manizales, ésta es una ciudad industrial comunicada por vía terrestre con Ibagué, desde allí se transportan vía terrestre alimentos no perecederos, bebidas lácteas y derivados de la leche.
- Las otras ciudades que conforman un segundo umbral de buenos modelos son en su orden: Armenia y Neiva. Armenia es productora de café, plátano y yuca, de allí provienen la mayor cantidad de estos productos. Productos de este mismo renglón provienen de Neiva, pero en menor escala.
- La distancia entre ciudades no es un factor que incida significativamente para este tipo de modelos. Esto se evidencia en la cercanía de Ibagué con Bogotá y con Armenia, no obstante los modelos no fueron los mejores.
- La metodología conocida como Datos Funcionales se sugiere para estudiar temas relacionados con el IPC. En ellos se recomienda tener una muy grande cantidad de datos, lo cual implica muchos periodos de tiempo.

5. Recomendaciones

- Se recomienda explorar nuevos modelos de regresión *spline* usando como covariables otros índices tales como: índice de desempleo, precio del dólar, etc., con el fin de encontrar modelos más robustos.
- Utilizar técnicas modernas de series de tiempo como modelos TAR, para la interpretación de este tipo de índices de caracter económico.
- Utilizar la moderna herramienta de Datos Funcionales, para la adecuada interpretación de modelos de carácter semiparamétrico, en los que el tiempo y el espacio juegan un papel importante.

Bibliografía

[1] Agresti, A. *Foundations of Linear and Generalized Linear Models*. Jhon Wiley & Sons Inc., New York, 2015.

[2] Aranda-Ordaz, F.J. (1981). On the two families of transformations to additivity for binary response data. *Biometrika*. **68**, 375-364.

[3] Atkinson, A.C. (1985). *Plots, Transformations and Regressions*. Oxford Statistical Sciences Series. Oxford.

[4] Belsley, D.A.; Kuh, E. and Welsch, R.E (1980). *Regression Diagnostics*. Second Edition. John Wiley, New York.

[5] Blanchard, O. (2008). The state of macro.*NBER*. Paper number 14259.

[6] Cook, R.D. (1986). Assessmen of local influence (with discussion). *Journal of the Royal Statistical Society*. B **48**, 133-169.

[7] Dobson, A. *An Introduction to Generalized Linear Models*. CRC Press Book, Third Edition, Sydney, 2008.

[8] Giraldo, A. and Zapata, C (2008). Probabilistic Model for the Phenomena of Transfer among Undergraduate Programs and Student Desertion. *Scientia Et Technica*. **39**.

[9] Gu, C. (2002). *Smoothing Spline ANOVA Models*. Springer, New York.

[10] Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*. Chapman & Hall, London.

[11] Jorgensen, B. (1987). Exponencial dispersion models (with discussion). *Journal of the Royal Statistical Society*. B **49**, 127-162.

[12] Kleinbaum, D. (1994). *Logistic Regression*. Springer-Verlag, New York.

[13] Lambert, D. (1992). Zero-inflated Poisson Regression with an Applications to Defects Manufacturing. *Technometrics*. Vol. 34, **1**, pp. 1-14.

[14] Lee, E.T. and Wang, J.W. (2003). *Statistical Methods for Survival Data Analysis*. John Wiley & Sons Inc., New York.

[15] Little, R.J.A. and Rubin, D.B. (1987). *Statistical Analysis with Missing Data*. John Wiley & Sons Inc., New York.

[16] Nelder, J.A, and McCullagh, P. *Generalized Linear Models*. Chapman & Hall, London, 1989.

[17] Nelder, J.A, and Wedderburn, R.W.M *Generalized Linear Models*. Journal of the Royal Statistical Society. Series A (General), Vol. 135, No. 3, pp. 370-384, 1972.

[18] Onu, B.M. (2006). *Manual de índice de precios al consumidor, Teoría y Práctica* . Oficina Internacional del Trabajo; Fondo Monetario Internacional; Organización de Cooperación y Desarrollo Económicos; Oficina Estadística de las Comunidades Europeas; Organización de las Naciones Unidas; Banco Mundial.

[19] Paula, G.A. *Modelos de Regressão com apoio computacional*. Versões 2004 é 2013. Universidad de São Paulo, S.P. 2013.

[20] Rao, C.R. (1973). *Linear Statistical Inference and Its Applications*. Second Edition. Wiley, New York.

[21] Sen, P.K. and Singer, J.M. (1993). *Large Sample Methods in Statistics: An Introduction with Applications*. Chapman & Hall, London.

[22] Vallecilla, J. (2011). *Café y crecimiento económico regional: El Antiguo Caldas 1870-1970*. Universidad de Caldas, Manizales.

[23] Viloria, J. (2007). *Economía del departamento de Nariño: Ruralidad y aislamiento geográfico*. Documentos de trabajo sobre economía regional. Banco de la República, Cartagena.

[24] Wahba, G. (1990). Spline Models for observational data. *Philadelphia: SIAM*.

[25] Wood, S.N. (2000). *Generalized Additive Models: an introduction with R*. Chapman & Hall/CRC Texts in Statistical Science, London.

Universidad del Tolima	PROCEDIMIENTO DE FORMACIÓN DE USUARIOS AUTORIZACIÓN DE PUBLICACIÓN EN EL REPOSITORIO INSTITUCIONAL	Página **1** de **3** Código: GB-P04-F03 Versión: 03 Fecha Aprobación: 15 de Febrero de 2017

Los suscritos:

JORGE EDGAR SILVA VELOZA	con C.C N°	93.376.140 de Ibagué
	con C.C N°	
	con C.C N°	
	con C.C N°	
	con C.C N°	

Manifiesto (an) la voluntad de:

Autorizar [X]

No Autorizar [ ] **Motivo**: _____

La consulta en físico y la virtualización de **mi OBRA**, con el fin de incluirlo en el repositorio institucional de la Universidad del Tolima. Esta autorización se hace sin ánimo de lucro, con fines académicos y no implica una cesión de derechos patrimoniales de autor.

Manifestamos que se trata de una OBRA original y como de la autoría de LA OBRA y en relación a la misma, declara que la UNIVERSIDAD DEL TOLIMA, se encuentra, en todo caso, libre de todo tipo de responsabilidad, sea civil, administrativa o penal (incluido el reclamo por plagio).

Por su parte la UNIVERSIDAD DEL TOLIMA se compromete a imponer las medidas necesarias que garanticen la conservación y custodia de la obra tanto en espacios físico como virtual, ajustándose para dicho fin a las normas fijadas en el Reglamento de Propiedad Intelectual de la Universidad, en la Ley 23 de 1982 y demás normas concordantes.

La publicación de:

Trabajo de grado	[X]	Artículo	[ ]	Proyecto de Investigación	[ ]
Libro	[ ]	Parte de libro	[ ]	Documento de conferencia	[ ]
Patente	[ ]	Informe técnico	[ ]		

Otro: (fotografía, mapa, radiografía, película, video, entre otros) [ ]

Producto de la actividad académica/científica/cultural en la Universidad del Tolima, para que con fines académicos e investigativos, muestre al mundo la producción intelectual de la Universidad del

Tolima. Con todo, en mi condición de autor me reservo los derechos morales de la obra antes citada con arreglo al artículo 30 de la Ley 23 de 1982. En concordancia suscribo este documento en el momento mismo que hago entrega del trabajo final a la Biblioteca Rafael Parga Cortes de la Universidad del Tolima.

De conformidad con lo establecido en la Ley 23 de 1982 en los artículos 30 ***"...Derechos Morales. El autor tendrá sobre su obra un derecho perpetuo, inalienable e irrenunciable"*** **y 37 *"...Es lícita la reproducción por cualquier medio, de una obra literaria o científica, ordenada u obtenida por el interesado en un solo ejemplar para su uso privado y sin fines de lucro".* El artículo 11 de la Decisión Andina 351 de 1993, "*los derechos morales sobre el trabajo son propiedad de los autores*" y en su artículo 61 de la Constitución Política de Colombia.**

- Identificación del documento:

Título completo: ESTUDIO DEL IPC EN EL DEPARTAMENTO DEL TOLIMA Y DEPARTAMENTOS ALEDAÑOS MEDIANTE GAMs

- Trabajo de grado presentado para optar al título de:

Especialista en Estadística

- Proyecto de Investigación correspondiente al Programa (No diligenciar si es opción de grado "Trabajo de Grado"): _____

- Informe Técnico correspondiente al Programa (No diligenciar si es opción de grado "Trabajo de Grado"): _____

- Artículo publicado en revista: _____

- Capítulo publicado en libro: _____

- Conferencia a la que se presentó: _____

Quienes a continuación autentican con su firma la autorización para la digitalización e inclusión en el repositorio digital de la Universidad del Tolima, el:

Día: **10** Mes: **Julio** Año: **2018**

Autores:		Firma		
Nombre:	JORGE EDGAR SILVA VELOZA		C.C.	93.376.140 de Ibagué
Nombre:			C.C.	
Nombre:			C.C.	
Nombre:			C.C.	

El autor y/o autores certifican que conocen las derivadas jurídicas que se generan en aplicación de los principios del derecho de autor.